EAI Endorsed Transactions
on Collaborative Computing

Research Article ICST.ORG

Tracing Coordination and Cooperation Structures via Semantic Burst Detection

Yu-Ru Lin[1,*], Drew Margolin[2], David Lazer[3]

[1]School of Information Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA
[2]Department of Communication, Cornell University, Ithaca, NY 14850, USA
[3]Political Science Department, Northeastern University, Boston, MA 02115, USA

Abstract

Developing technologies that support collaboration requires understanding how knowledge and expertise are shared and distributed among community members. We explore two forms of knowledge distribution structures, coordination and cooperation, that are central to successful collaboration. We propose a novel method for detecting the coordination of strategic communication among members of political communities. Our method identifies a "rapid semantic convergence," a sudden burst in the use linguistic constructions by multiple individuals within a short time, as a signature of coordination. We apply our method to the public statements of U.S. Senators in the 112th U.S. Congress and construct coordination and cooperation networks among these individuals. We then compare aspects of these networks to other known properties of the Senators. Results indicate that the detected networks reflect underlying tendencies in the social relationships among Senators and reveal interesting differences in how the different parties coordinate communication.

Received on 17 July 2014, accepted on 23 July 2014, published on 20 October 2014

Keywords: semantic burst, semantic convergence, burst detection, coordination, cooperation, social networks, public statement, political network, strategic communications

doi: 10.4108/ cc.1.2.e7

1. Introduction

Developing technologies that support collaboration requires an understanding of how knowledge and expertise are shared and distributed among individuals in both formal organizations or more informal social groups. However, knowledge distribution structures vary greatly with the culture and inner workings of different groups. These endogenous structures implicitly influence how group members interact with each other and perform as a whole. Hence, capturing the knowledge distribution structures specific to individual groups has been an intriguing problem in studying human collaborations. In this article, we explore two forms of knowledge distribution structures, coordination and cooperation, that are central to successful collaboration, among members in political communities.

According to Engesröm, coordination and cooperation are two of the fundamental forms in human interaction[1] [1]. At the level of coordination, each actors work independently without explicit communicating with each other, while in cooperation, actors try to find mutually acceptable ways to solve a shared problem [1]. In this article, we propose a novel method for detecting the structures of coordination and cooperation from communication data automatically.

An obvious challenge in this research is the difficulty of obtaining data and assessing the results. The recent increase in the availability of enormous digital archives of communication behavior offers a novel opportunity to address this issue [2]. Here, we utilize the public statements by U.S. Senators in our research. While publicly available data do not directly reveal coordination and cooperation structures, patterns in these trace data can suggest when coordination by some mechanism appears to be operating [3, 4]. By identifying these cases, analysis can then be tuned more finely to examine the potential causes of and processes involved in this coordination.

In the context of politics, communicating effectively through public statements is important for politicians. Through effective strategic communication, politicians can influence both media and voters, promoting attention to favored positions as well as favorable interpretations for their own policies and unfavorable interpretations of opponents and their views [5–7].

Despite the recognition that strategic communication is important to a politician's ability to gain power

Corresponding author. Email: yurulin@pitt.edu
*[1]The third form is reflective communication.

and win elections, little research has considered the social processes that influence politicians' communication strategies. Outside of the study of politicians and those in power, a variety of research suggests that coordination and cooperation in strategic communication are critical to the success and failure of political and social movements [8, 9]. While it is possible that the achievement of formally elected positions of power reduces or obviates the need for strategic communication coordination, there is also reason to expect this would not be the case. At the very least, political parties appear to be highly influential in the persuasion of audiences [10, 11]. Furthermore, many of the arguments for the advantages that elected politicians possess in strategic communication, such as the ability to provide information subsidies to media outlets, suggest that pooling and coordinating resources across individuals would also have benefits [12, 13].

In this study we focus on one such pattern of traces: rapid semantic convergence, which we defined as sudden bursts in the frequency with which particular phrases, measured as trigrams, are used in the public statements of U.S. Senators. The basic logic of this approach is that rapid semantic convergence indicates a coordinating mechanism – a causal process that brings senators' together linguistically.

We articulate four broad candidates for these coordinating processes: emergent contexts and events, shared persuasive interests, rhetorical innovation, and collaboration. For each of these categories we describe the extent to which the process that leads to convergence is a matter of individual and/or collective choices on the part of the individuals. If the behavioral structural signatures, in the form of shared language, can be identified, the incidence of these signatures may provide substantial insight into how politicians coordinate their activities and influence on another.

The article is structured as follows. Followed by a review of related work, we provide the theoretical and empirical foundations for our approach. We illustrate these theoretical processes with examples detected by our method from the public statements of U.S. Senators. We then articulate our approach for detecting rapid semantic convergence using two methods – a burst detection algorithm as well as a means for detecting joint authorship of public statements. Using these methods, we generate three networks built from the tendency for pairs of senators to suddenly deploy similar language. We compare the structures of these networks and explore their relationship to covariates, such as shared committee membership networks and party leadership structure.

2. Related Work

There are several ways to measure the underlying construct of "rapid semantic convergence." We briefly review three principle methods for doing so and describe their strengths and weaknesses in capturing the phenomenon of interest.

Correlations in Semantic Frequency. A fundamental question in semantic analysis is semantic representation and extraction. Semantic representation deals with the problem concerning the relationship between "concepts" and "word meanings." Popular approaches include semantic networks and co-occurrence models. Semantic networks is a network based representation that represents the meaning of each word by its relation of other words. For example, in WordNet [14], words are represented as nodes and semantic relationship are labelled connections between them. It is based on holistic views [15] which assumes a non-decomposable, one-to-one mapping between the lexical representation (the word) and conceptual representations of things, events, etc. In such representation, the connections between words are constructed based on prior knowledge.

A different approach based on co-occurrence analysis seeks to learn the representations of words in terms of their relationship to other words, automatically from corpora of texts. It is based on the assumption that similar words tend to appear in similar contexts. The approach can be found in widely-adopted vector space models such as Latent Semantic Analysis (LSA) [16] and probabilistic models such as Probabilistic Latent Semantic Analysis (pLSA) [17] and Latent Dirichlet Allocation (LDA) [18].

In Latent Semantic Analysis (LSA) [16], a document is represented as a vector where each dimension corresponds to a separate feature (a term) from the document. The entire corpus is represented as a term-document matrix and the values are commonly determined by the tf-idf weighting scheme [19]. The idea of LSA is to project the documents and their term features into a lower-dimensional latent concept space in order to represent a relation between the terms and some concepts, and a relation between those concepts and the documents. The low-dimension semantic latent space is obtained by decomposing the term-document matrix using Singular Value Decomposition. Despite its success for modeling implicit semantic structures between documents and words, one issue with this approach is that the resulting dimensions might be difficult to interpret, for example, the LSA approximation of the term-document matrix may contain negative values.

The Probabilistic Latent Semantic Analysis (pLSA, or the aspect model) [17] was introduced to overcome the weakness of LSA. It is based on a generative model that associates a latent variable with each occurrence

of a word in a document. The Latent Dirichlet Allocation (LDA) [18] further improves the pLSA by introducing a Dirichlet prior on document-topic distribution. LDA represent documents as mixtures of topics (like "concepts" in LSA or "aspects" in pLSA), where a topic is a probabilistic distribution over words. Compared to the LSA model, the probabilistic models provide ways of interpreting the relationships between document-topic and topic-word in terms of probability weights.

Such topic mixing representation effectively is compact (the number of topics is significantly fewer than the number of terms) while still preserving salient statistical relationships. However, the resulting topics are synthetic and do not explicitly correspond to the prior knowledge of document topics. Furthermore, the meaning of words are determined without considering its specific contextual use in the documents, which makes it difficult to inform a potential social process corresponding to the particular use of words.

Burst Detection. In time series data, the presence of a burst suggests that the occurrence of a data feature or value is unexpectedly frequent in a short period. This unexpected occurrence is often associated with an unusual event. Intuitively, burst detection can be achieved by identifying a burst region where the data value exceeds certain threshold. The threshold can be determined based on heuristics [20], different data distribution assumptions [21] or statistical tests [22]. The Cumulative Sum (CUSUM) method [23] is one of the most popular statistical approach for change point detection. However, threshold-based methods lack flexibility to recognize bursts with various lengths, for example, a longer burst may be identified as several short bursts. Kleinberg [24] proposed a state-based model using Hidden Markov Model (HMM), which extends the threshold-based method with a more relaxed threshold. The idea behind this method is that it models the state transitions as low-probability events, and a cost function is assigned such that a smooth state tends to be more persistent than transitions. Mane and Böner [25] used this method to track the temporal evolution of major topics in scientific publication. In their study, the potential topic words are pre-specified according to Biologists' domain knowledge. [2]. Although the state-based burst detection method has the ability to identify longer bursts with noises, it remains a challenge to deal with emergent topics or ambiguous signals – the variable notion of threshold makes it difficult to recognize whether there is a drift in the meaning of a given term.

[2]The topics are selected based on Institute for Scientific Information (ISI) key words and MEDLINE's controlled vocabulary (MeSH terms)

As described above, previous work has focused on identifying the meaning of terms and expressions observed in semantic trace data and grouping documents and individuals based on these shared meanings. In these contexts, rapid semantic convergence would represent shared understandings and intentions among group members [26, 27]. Yet another way to conceive of rapid semantic convergence is as the residue of a group process that imposes itself on individual behavior independent of textual meaning. More precisely, when individuals converge in their public use of language it might be because, as typical models assume, they have reached a consensual agreement regarding a shared set of ideas, with their words reflecting this unified psychological state. However, their convergent semantic behavior may also reflect the fact that the group, or incentives within the group, has the power to encourage them to issue statements that differ from their personal views or which mean things they do not personally intend or even technically understand. That is, incentives for conformity or specialization may lead group members to parrot one another's language independent of the meaning of the phrases or their precise feelings about them.

The possibility that the higher order structures and incentives of the group as a whole encourage or compel semantic convergence suggests it may be useful to analyze semantic convergence in a different way. In particular, it may be useful to assess the observable features of individuals and the social relations that lead them to converge with one another independent of the meaning of the statements around which such convergence takes place. Thus, in the next section, we enumerate several of these higher order incentives.

3. What Leads to Semantic Bursts?

In this section, we describe theories on the sources of semantic convergence.

Shared Categories for Emergent Features and Events. A fundamental property of language as a communication tool is its use of shared symbols to refer to particular referents [28]. Thus, a basic reason why individuals may converge in their use of words or concepts is that they are responding to experiences of referents that they share in common. For example, these referents might be features of the environment or events that have taken place [29].

When applying concepts to describe situations or identify particular ideas or entities, individuals tend to begin by applying basic categories [29, 30]. Basic categories are those that best balance the trade-off between specificity in communicating information and availability to speakers and familiarity to audiences [30, 31]. That is, the concept is specific enough to carry a narrow set of appropriate applications but

general enough so that it can be applied frequently, making it accessible through memory and familiar to different individuals. Basic categories settle in particular locations based on the frequency with which different features appear in the environment [30, 32]. See [29] for a review. For example, children tend to learn the category "bird" before learning more specific categories such as "robin" [32]. Once experienced in identifying "birds," children then move on to make finer distinctions based on more commonly experienced kinds of birds (e.g. robins) [32, 33]. Experts tend to share more specific categories, such that individual items may often have highly specific names [34–36].

Basic categories help make communication intelligible to an audience. The incentive to conform to basic category use is thus often a response to the state or structure of an audience, even when that audience is unseen. [37] show that in communities with a cohesive social structure, individuals are more likely to articulate statements in commonly held terms. When social structure is more fractured, they revert to more idiosyncratic expressions that may more precisely reflect their own ideas but are not well understood by others.

In order to gain or maintain visibility and influence with media and constituents, politicians may be compelled at times to respond to the news of the day or novel events [6, 38]. These events may cause semantic convergence by compelling politicians to describe or take a position for which the set of basic categories or specialized terms is already convergent [39].

For example, the death of Osama bin Laden in May, 2011 had important implications for both U.S. foreign and domestic policy. Accordingly, several U.S. Senators issued press releases immediately following the report of his death[3]. Furthermore, since bin Laden's name was widely known and recognized, Senators referred to him by name. Thus, our method reveals a sudden convergence in the use of the name "Osama bin Laden" in Senators public statements. Ninety senators used the name "Osama bin Laden" in a public statement within approximately 1 week of his death.

Politicians also must address more mundane, scheduled events or changes in policy context. As the agenda for debate and discussion shifts from one proposed policy to another, the categories and entities that politicians name will also shift accordingly [40]. The beginning of debate on a particular topic within the Senate can lead several Senators to comment publicly on it, leading to semantic convergence. This convergence is due to the limited set of categories that can be used to describe aspects of the topic in a way that is broadly intelligible to the public and the media. For example, in early October, 2011, the Senate debated the terms of trade with several nations. The basic category used to refer to these contracts is "free trade agreements." Thus, the phrase "free trade agreements" was used by 33 Senators within a one week period[4]. While this may have been the result of strategic coordination (negotiation between the Senators in which bill to discuss), the semantic convergence it breeds may not be the result of any explicit cooperation between senators that discuss it using the same terms. Thus, cases such as these represent exogenous sources of strategic communication coordination.

Shared Interests in Persuasion. Political discourse is inherently persuasive. Politicians seek electoral advantage in their use of public statements [7, 40]. Politicians can influence the ways in which their constituents will interpret their actions by identifying information, arguments, and frames which justify their point of view and cast their decisions in the most favorable light [5].

Although politicians' electoral fates are ultimately individual, the cooperative nature of political action and the strength of political parties in influencing election outcomes lead politicians to share a variety of persuasive interests. Legislators work together to craft legislation [41]. Co-sponsors of a bill thus share a persuasive interest in the public interpreting the bill and the reasoning behind it in a positive light. More broadly, public interpretation of politicians' individual positions is strongly influenced by party identification and the favorability of the party [10, 11]. Thus, members of the same party share an interest in framing and justifying policies in a manner that benefits their party.

Shared persuasive interest can lead to semantic convergence in two different ways. First, for some individuals, a credible position must be supported by specific reasons and evidence [42]. To persuade these individuals to adopt a particular point of view, politicians will likely draw on specific sources or pieces of evidence. When a particular topic is debated, politicians sharing persuasive interest will exhibit semantic convergence by virtue of their citation and invocation of common arguments, evidence and sources.

Though citizens may often ignore specific information and evidence and rely instead on peripheral cues or sources, media outlets may be more susceptible to

[3]For example, see http://www.baucus.senate.gov/?p=press_release&id=459, http://durbin.senate.gov/public/index.cfm/pressreleases?ID=1ddceb11-cde0-46bd-a7fd-83513f5b80b9, http://www.mikulski.senate.gov/media/pressrelease/5-2-11.cfm,http://boxer.senate.gov/en/press/releases/050211.cfm

[4]For example, see http://conrad.senate.gov/pressroom/record.cfm?id=334460, http://ronjohnson.senate.gov/public/index.cfm/press-releases?ID=534b1e62-4f9d-418b-a654-6c627a08825c, http://rockefeller.senate.gov/press/record.cfm?id=334453

these techniques. Many newspapers and other media outlets have limited budgets for information gathering about national policy. Thus, these outlets often rely heavily on politicians to provide not only their positions but the facts and reasons which explain the issues relevant to the policy in question [13]. Since it is in the interest of the politician to provide background information which casts their own position in a favorable light, biased information often becomes the basis of the news that is reported [12]. Influencing these outlets can benefit politicians as voters often cannot identify or remember the sources from which they have received political information [43]. Other media outlets are explicitly biased and seek information to justify their pre-determined support for a particular view [44]. This effect should lead to substantial semantic convergence due to shared persuasive interest as the gains that politicians can achieve by disseminating a new fact that supports their position may be substantial [7].

For example, in January 2012 several Republican Senators advocated for the approval of the Keystone XL pipeline project. In justifying this position, several of them referred to statistics regarding the increase in oil production the pipeline was expected to yield. During an 8 day period toward the end of the month, 24 Senators used the phrase "oil per day" in their public statements, including several as part of a joint press release[5]. In another example, Senators often cite scores and analyses issued by the Congressional Budget Office to bolster their justification for or critique of a particular bill. On several occasions, the number of Senators referring to the "Congressional Budget Office" increased to more than 20 individuals within a short time period.

In these examples, semantic convergence occurs due to a shared interest in persuasion and an economy in research costs achieved by repeating the same justifications. Much as politicians subsidize the media, they may also subsidize one another, particularly if the knowledge they provide is easy to copy or imitate. One senator may find a useful statistic or report and cite it, revealing its relevance to others. There is also evidence of cooperative dissemination of evidence and arguments through the release of "talking points" from think tanks and other partisan or issue based interests [45].

A second way that shared persuasive interest can lead to semantic convergence is through the use of frames [5, 46]. Frames are deployed to suggest interpretations for particular events or policies by highlighting certain aspects of a situation and suppressing others. While for many entities and ideas there are clearly identifiable, unique basic categories or specific names, for others there may be a distribution of appropriate terms or phrases with similar yet imperfect fit [31]. There may exist no dominant term which captures all of the relevant features of a situation, and thus permitting politicians to choose from a set of candidate phrases that frame the issues in a way that is beneficial to them [29, 47].

Frames appear to be particularly important in persuading individuals that do not pay close attention to specific arguments [38]. An individual may find both of two contradictory frames to be resonant and persuasive [48]. This suggests that when persuading through framing, only the effectiveness of the chosen frame need to be considered. This leads to "cross-talk" in political campaigns in which opponents use different words to discuss the same issues and rarely acknowledge one another's frame, eschewing the other's favored terms [7]. Repetition is also important for frame resonance [8]. The more experience an audience has with a frame, the more credible or legitimate this frame becomes to that audience. Similarly, as a greater number of individuals use a frame, it gains in legitimacy and credibility [49].

These factors suggest that politicians that share a persuasive interest will benefit substantially from a strategically coordinating their use of frames. A frame which is not optimal for a particular politician's individual agenda may nonetheless be effective if others, such as those in his or her party, agree to use it as well. This suggests that politicians may negotiate to use a consistent set of frames.

For example, at the beginning of the 112th Congress Republican Senators appeared to attempt to consolidate support for their party and opposition to the President. In particular, they highlighted what they perceived to be the failures of the Obama administration through its first mid-term election. In a four day period, ten Republican Senators used the phrase "two years ago" to refer to different aspects of the President's failed policies, including the stimulus package and foreign policy initiatives.

Rhetorical Innovations. The most supportive arguments and facts and the most effective frames are not always obvious or easy to discover. Most individuals tend to possess only a limited number of the total pool of arguments that they would find persuasive on a topic [50]. The supply of frames may be limited as well. Since satisfactory frames are generally sufficient to be persuasive, there is limited need to explore a large set of alternatives to construct a persuasive message [48, 51]. Also, the resonance of novel frames is often difficult to observe without first witnessing the reaction of an audience [52, 53]. These factors create an incentive for

[5]For example, see http://www.lee.senate.gov/public/index.cfm/2012/1/bipartisan-group-of-senators-to-introduce-legislation-to-approve-keystone-xl-pipeline

politicians to let others invent and try a novel frame before they adopt it themselves.

As a result, politicians may often coordinate their use of words without explicitly conferring or agreeing to do so. This can lead to a diffusion process for rhetorical innovations [54]. That is, there may be limited investment in the development of new phrases or optimization of frames [27], but if and when one politician happens upon a useful name, fact, or word combination, others that are exposed to the new combination may quickly adopt it and use it in their own statements. This is particularly likely for short, easy to imitate phrases with limited complexity in appropriate use [55]. Like a disease that, once it has "infected" one host can easily exploit other hosts to which it gains access, the novel concept or combination may be latent for some time and suddenly burst through a population of politicians [56, 57]. Thus, it can be expected that a certain degree of discursive imitation of novel phrases or conceptual combinations will occur, leading to semantic convergence with the onset of an "infection" of one individual [56, 58, 59].

Furthermore, this imitation may be enhanced by the increased recognition and legitimacy the phrase will obtain as it is used more frequently. This can lead not only to increased use of the phrase but also to the re-application of the phrase to new contexts [60]. That is, once the phrase has been recognized as effective in a particular context, it may be used so frequently as to be more broadly recognizable. Individuals may then import the phrase into new contexts.

For example, following the death of Osama bin Laden, several senators attempted to convey that bin Laden's death, while an important step forward in the effort to limit terrorism, was not the sign that the task was complete. Within 7 days of bin Laden's death, 23 of the Senators that issued statements on this topic used the phrase "must remain vigilant" to implore the continued commitment to anti-terror efforts. This example will be described in more details in the result section.

Teamwork and Collaboration. The preceding processes for semantic convergence are largely built from individual incentives and habits. For example, Senators would not need to explicitly work together or choose to speak similarly for them to rely on the same basic categories or for one Senator to copy the arguments or imitate the rhetorical innovations of another. At the other end of the spectrum is the purposeful teamwork and collaboration. When individuals work together, they must negotiate a shared language which facilitates communication about the topics on which they are working [61]. Having identified a shared language and way of framing a situation, it is easier for others to adopt the same terminology.

In these cases, the terms used are explicitly agreed to by all members of the group [62, 63]. The joint authorship of public statements is in this way similar to co-sponsorship of bills [41]. The time required to develop and negotiate a jointly acceptable statement should be substantially less than that required to develop and negotiate jointly acceptable legislation, however. The two may also overlap, as politicians agree to collaborate in their communication to support their collaborative work on legislation.

For example, in February, 2012, 14 Democratic Senators wrote a letter to the Senate to pass a "payroll tax cut." This letter was then followed by 35 other Senators (mostly Democrats) using this phrase in their own public statements. In this case, explicit collaboration and the semantic burst it produced appeared to push the term onto the agenda for others to consider.

Summary. We are interested in identifying relationships of strategic communication coordination via the observation of rapid semantic convergence. In this section we have reviewed theoretical arguments that suggest that rapid semantic convergence can be an indicator of higher level processes influenced by needs for coordination and collaboration. In the next section, we begin by demonstrating how our algorithm detects semantic convergence. We also present a method for detecting explicitly negotiated convergent communication in the form of jointly authored public statements.

Based on these patterns of convergence, we then build network structures showing the underlying coordination relationships between individual senators. We then explore how these explicit collaborations and implicit coordination and diffusion relationships correspond to a other measures of Senator attributes, positions, and network structures.

4. Method

4.1. Data

As of March 2012 we have gathered 0.4 million documents from the public statements of Members of the US Congress from the Vote Smart Project website[6]. Fig. 1 shows the number of public statement documents gathered in our dataset. According to Vote Smart, the 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 official[7]. In this study we focus on the statements made

[6] http://www.votesmart.org

[7] In future analyses we will disaggregate the analysis by type of public statement.

by the members of the 112th Senate, during the period between January 2011 and March 2012. We retrieve the individual attributes for the Members of Congress using Sunlight Congress API[8].

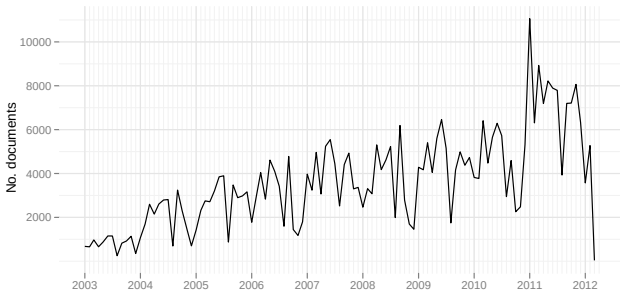

Figure 1. Monthly volume of public statements gathered in our dataset.

4.2. Semantic Burst Detection

Our goal is to identify a set of instances from the public statements of the Members of Congress where the use of certain words is shared among a group of members more frequently then usual and the use is concentrated within a short period of time. Such instances of semantic convergence are indicative of unobserved processes among the members that potentially influence the product of semantic convergence (in the observed public statements). We call an instance of bursty use of words "infection instance." For each of the identified infection instances, we can derive certain social relationships within the infected population. The overall procedure include three steps: (1) burst n-gram detection, (2) n-gram infection instance extraction, and (3) social network construction. We describe each step in the following.

We first parse the corpus of public statements to construct a term-document sparse matrix, where an entry (i, j) indicates the number of occurrences of term i in document j. The terms are n-grams with highest tf-idf weights. An n-gram is a contiguous sequence of n words from a given sequence of text. We use trigram ($n = 3$) in this paper. The presented analysis can be well extended to shorter or longer n-grams. Here we use trigrams because a trigram conveys more specific meaning than single word or bigram, but it has a greater advantage in terms of computational efficiency, compared with longer n-grams [64]. (Throughout this article, "n-gram", "term" and "word" are used interchangeably.) N-grams that contain stop words such as "the", "is", etc. are removed.

[8] http://services.sunlightlabs.com/docs/Sunlight_Congress_API/

For each non-zero entry in the term-document, we retrieve a 4-tuple (a, d, w, t) from the document metadata to represent the occurrences of a word w (n-gram) in a document d (public statement) given by an actor (i.e., a Member of Congress) at the time t. The time resolution is one day, consistent with the resolution obtained from the document metadata.

To detect a bursty use of an n-gram within a short period of time, we first construct a time series $w_i(t)$ for each n-gram w_i, where: $w_i(t)$ is given by number of actors who use the n-gram at least once at the time t.

We then use the following on-line filtered derivative algorithm to detect bursts within each $w_i(t)$ sequence. In this method, a change in the mean level of a sequence of observations is locally characterized by a great absolute value of the derivative of the sample observations [23]. Since the derivative operator may be sensitive to noises, a filtering operation is applied before derivation. Specifically, we consider the discrete derivative of f_k:

$$\nabla f_k = f_k - f_{k-1}, \tag{1}$$

where f_k is the decision function based on log-likelihood ratio test:

$$f_k = \sum_{i=0}^{N-1} \gamma_i \ln \frac{p_{\theta_1}(y_{k-i})}{p_{\theta_0}(y_{k-i})}, \tag{2}$$

And the burst alarm is activated at t_a if:

$$t_a = \min\{k : \sum_{i=0}^{N-1} \delta(\nabla f_{k-i} \geq h) \geq \eta\}, \tag{3}$$

where N is a fixed sample size, $\delta(x)$ is the indicator of event x, h is the threshold for the derivative, and η is a threshold for the number of crossings of h, which is usually used for decreasing the number of alarms in the neighborhood of the change due to the smoothing operation.

In the case of an increase in the mean, the decision function f_k corresponding to Eqn. 2 is:

$$f_k = \sum_{i=0}^{N-1} \gamma_i (y_{k-i} - \mu_0). \tag{4}$$

We choose an integrating filter with N constant weights γ_i, so the decision of alarm time is based on local averages of sample values.

When an alarm is activated for an n-gram w, we call the pair (w, t) a candidate n-gram infection, where $t = t_a$ is the onset time of the burst region. We shall extract a set of n-gram infection instances from the candidate set. For each candidate infection, we expand the time window on both sides to $[t_s, t_e]$ where $t_s < t < t_e$ and retrieve all actors who used the n-gram w at least once during the time window $[t_s, t_e]$. If an actor uses the

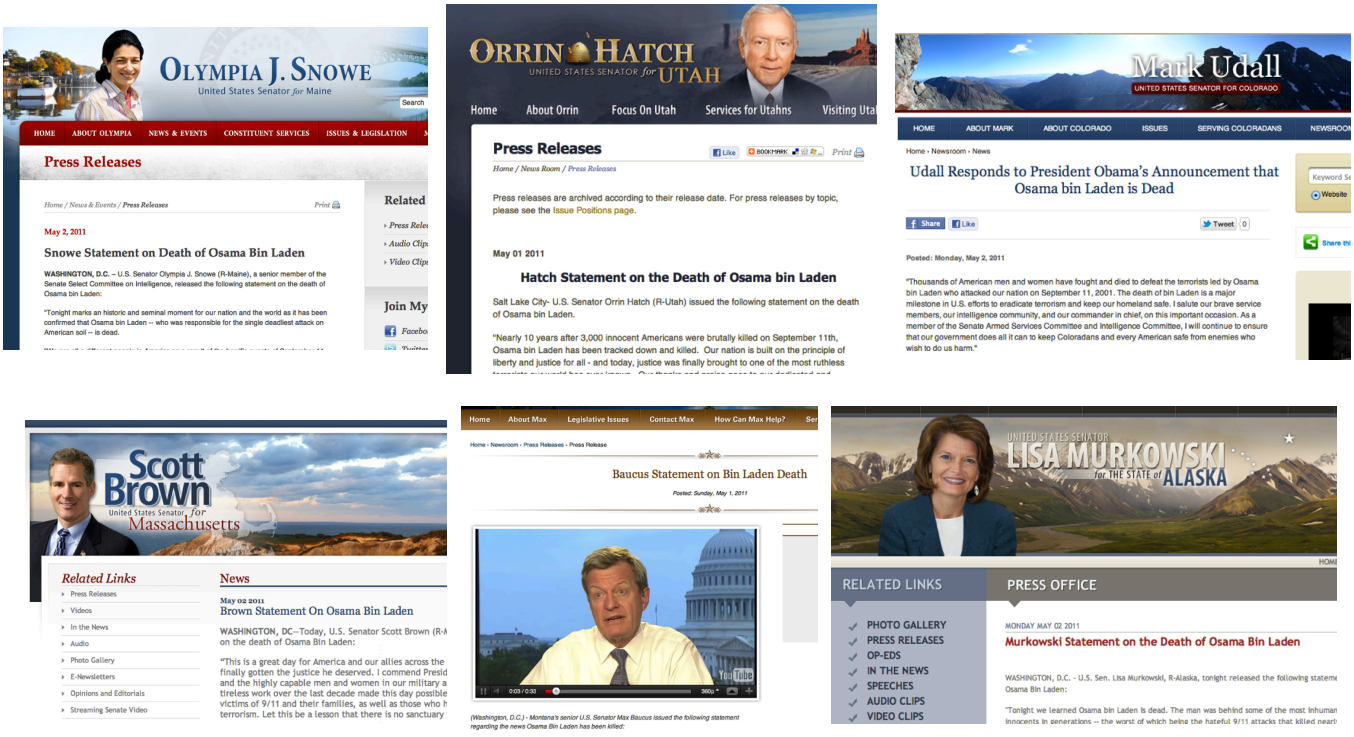

Figure 2. The trigram "osama bin laden" on Senators' web press release on May 2, 2011.

n-gram at different time points, only the first time is retrieved. An infection instance of (w, t) associated with time window $[t_s, t_e]$ is defined by the list of infected actors: $\{(a_1, t_1), (a_2, t_2), ...\}$, where $t_1 \leq t_2 \leq ...$ and $a_i < a_j$ if $t_i = t_j$.

If two candidate pairs have the same n-gram and different but overlapping infected durations, the pairs are merged to have a extended duration covering both durations. If two n-grams have the same onset time and infection instance, the two are called exchangeable n-grams and the infection instances are merged into a single instance. In other words, an infection instance may be associated with multiple n-grams if all of them are exchangeable with each other.

4.3. Network Extraction

The next step is to derive social relationships from the detected n-gram infection instances. Two types of networks are obtained from the instances: infection following and infection sharing networks.

The infection following network is calculated on the assumption that the point in time at which the actor uses the n-gram for the first time is important in the unobserved social processes. The timing makes no difference in the infection sharing network. The infection following network is a normalized weighted directed network, defined as:

$$W^F = \sum_q \frac{\theta_{ij}^q}{\sum_{ij} \theta_{ij}^q} \tag{5}$$

where for each n-gram instance q, the edge weight between two actors a_i and a_j in the instance is given by:

$$\theta_{ij}^q = \frac{exp(-\Delta t_{ij}/r)}{N(t_i)}, \tag{6}$$

if with $t_i > t_j$, and $\theta_{ij}^q = 0$ otherwise. $\Delta t_{ij} = t_j - t_i$, r is the exponential decay rate and $N(t_i)$ is the number of actors in the instance who are infected before t_i. This weighted scheme assigns higher weights to actors who are infected closely in time, and are more likely to be infected before others.

A symmetric infection network is defined as:

$$W^S = \sum_q \theta_{ij}^q \tag{7}$$

with edge weight $\theta_{ij}^q = 1$ for all pairs of actor (a_i, a_j) in an instance q, and $\theta_{ij}^q = 0$ otherwise.

In addition to the two networks, we observe a definite structure in our dataset – occasionally senators release exactly the same public statements on the same day via joint press release. This structure can be recognized through duplication detection of documents. We thus

construct a third network to reflect this structure as a comparison. The joint press network is given by:

$$W^J = \sum_k \theta_{ij}^k \tag{8}$$

with edge weight $\theta_{ij}^k = 1$ for all pairs of actor (a_i, a_j) in a joint press release k, and $\theta_{ij}^k = 0$ otherwise.

4.4. Analysis and Controls

We compare the structures of these extracted networks with additional information gathered for the Senate.

Covariate Networks. We assemble three covariate networks that are likely to be associated with the inferred networks: a shared party membership network, a shared committee membership network, and an adjacent home states network.

Shared Party Membership. Senators from the same party share substantial persuasion interests. Thus, it is expected that members of the same party will be more likely to converge semantically through explicit collaboration as well as through imitation or a shared evaluation of arguments or frames as useful.

Shared Committee Membership. Senators on the same committee might be prone to semantic convergence for several reasons. First, by virtue of working on the same committee, these Senators will likely come into contact with one another more frequently [65]. Physical proximity is an important predictor of communication and collaborative team formation [66, 67]. It is also a means through which Senators may be exposed to one another's rhetorical innovations. Also, because committees are formed to address specific issues, Senators on the same committee may share persuasive interests.

A network of shared committee membership was constructed from the Congressional Directory for the 112 Congress. A bipartite network of Senators-Committees was built from the directories and then projected into a one mode, weighted Senator-Senator network where a link represented the number of committees on which two Senators served together.

Adjacent Home States. Senators who come from states that are geographically close to one another may also share persuasive interests and rhetorical exposure [68]. Many issues that are important or salient for citizens of one state may also be important or salient for members of nearby states. For example, the proposed Keystone XL pipeline affects voters in the central states in which the pipeline would be constructed differently than it effects voters in coastal states. The persuasiveness of an argument or frame may also vary with local cultures [69]. Exposure to issues and to the frames in which they are described or interpreted may also correspond to geographic proximity. The reach of major media outlets, such as newspapers and television stations corresponds to the geographic boundaries of media markets, many of which cross state boundaries and thus serve adjacent states [70].

A network of senators with geographically adjacent home states was constructed. A link between two senators in this network indicates that these senators' home states border one another. For example, Barbara Boxer (D-California) has a link to John McCain (R-Arizona). Senators from the same state were also given a link.

Covariate Attributes. For an exploratory analysis, we compare network centrality scores calculated from the inferred networks to several attributes of individual senators. All attribute data are taken from the Voteview database unless otherwise specified.

Year Entered Senate. This attribute captures the first year that a senator joined the Senate. For senators that have been elected or appointed in non-continuous terms (there are two in the 112th Congress), the first year is used. This measure indicates the extent of a senator's seniority and experience level.

Leadership. Senators in leadership positions may also be leaders in a party's public communication strategy. Senators in leadership positions – Majority Leader, Minority Leader, Majority Whip, Minority Whip, Party Conference Chairs, and President Pro Tempore were dummy coded as leaders. This resulted in 6 Senators being coded as "leaders" (Harry Reid, Mitch McConnell, Dick Durbin, John Kyl, Daniel Inouye, and John Thune).

Term End. Senators come up for election in staggered "classes." One third of U.S. Senators are up for re-election in 2012, one third in 2014, and one third in 2016. The Term End thus measures the number of years until the Senator will face re-election. Senators with a lower Term End score should be more attuned to the persuasive effect of their public statements on voters.

DW-Nominate. The DW-Nominate score indicates the extent to which a politician is ideologically conservative (liberals receive negative scores with large absolute values) [71]. Senators with more moderate scores may have more in common with a larger number of Senators, drawing from moderates in both parties. Senators with more extreme views may be more likely to identify arguments, evidence, or frames which support a particular point of view. These scores were taken from Voteview.com based on the calculations for the 111th Congress. Thus, senators that are new in the 112th congress were not included for this analysis unless the site provided a score based on a senators record in the House of Representatives during the 111th Congress.

5. Results

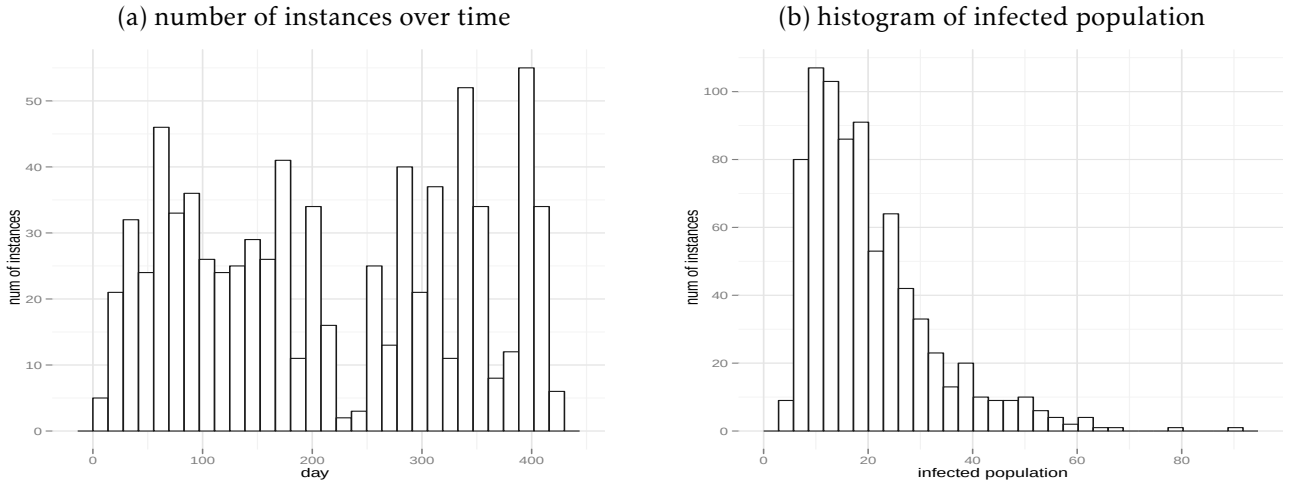

Figure 3. (a) Number of n-gram infected instances extracted by our method. (b) Histogram of infected population per instance. The n-gram with the largest population is "osama-bin-laden" which infects 91 members.

5.1. Detected Semantic Bursts

In this section we report results of analyses performed in the networks inferred by our method. Our method detected 783 bursts over the 426 day observation period (14 months). In Fig. 3 we show the number of detected semantic bursts (infected instances) over time and their significance in terms of individuals involved in the burst instances.

Figure 4 shows the bursts used as examples in the theory section of this article. Each chart shows the daily usage of the identified trigram over the observation period. The burst periods are depicted in red. These period are identified as the high peak period and extended to seven days before and after this period.

These figures show that the detection method identifies clear bursts. In each case, the peak of the trigram usage is substantially larger than its typical usage rate. The bursts also show varying degrees of decay. These patterns may suggest manners in which the kinds of bursts may be distinguished via these methods.

Figure 4(a) and (e) show the bursts for basic categories and named entities: "Osama bin Laden" and "free trade agreements." These are examples where external events, rather than negotiated or other less explcit coordination processes are responsible for the burst. In the case of Osama bin Laden, the absolute mean increase in infected population is 5.91 (with a 11-day average smoothing window), which means there are more than 65 more senators who used the n-gram "osama bin laden" as compared to the number of senators using the same n-gram before the onset time of the infection. There are in total 91 Senators that used this name during the week before and after the onset time (Fig 4(a) red period). For "free trade agreements,"

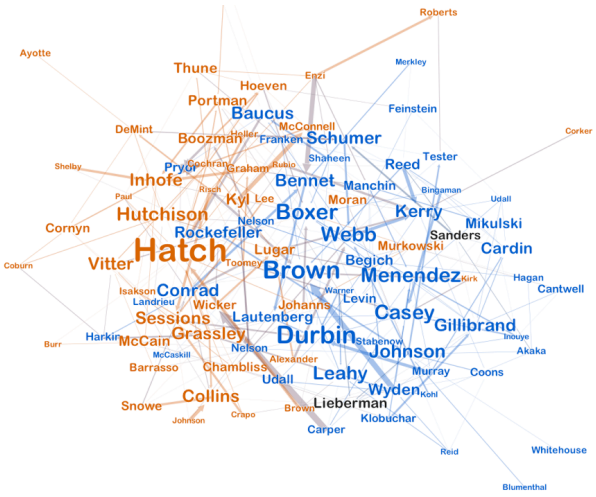

Figure 6. Infection Following Network: the network of who follows whom in public statements using bursting trigrams.

the use of this category has the absolute mean increase 3.73.

Figure 4(b) shows the burst for "oil per day," a reference to a statistic regarding the projected yield from the construction of the Keystone XL pipeline. This burst is smaller in magnitude (absolute mean increase 1.45). Nonetheless, it is still clear in the picture that this is a substantial increase. The use of this phrase also appears to show a modest diffusion pattern. Once the window is closed, this phrase appears to be used more frequently than in the period prior to the window, though still at a much lower rate than during the burst itself.

Figure 4(c) shows a similar pattern for the burst in the use of the frame "two years ago." The use of this frame has absolute mean increase 1.0. This phrase was used

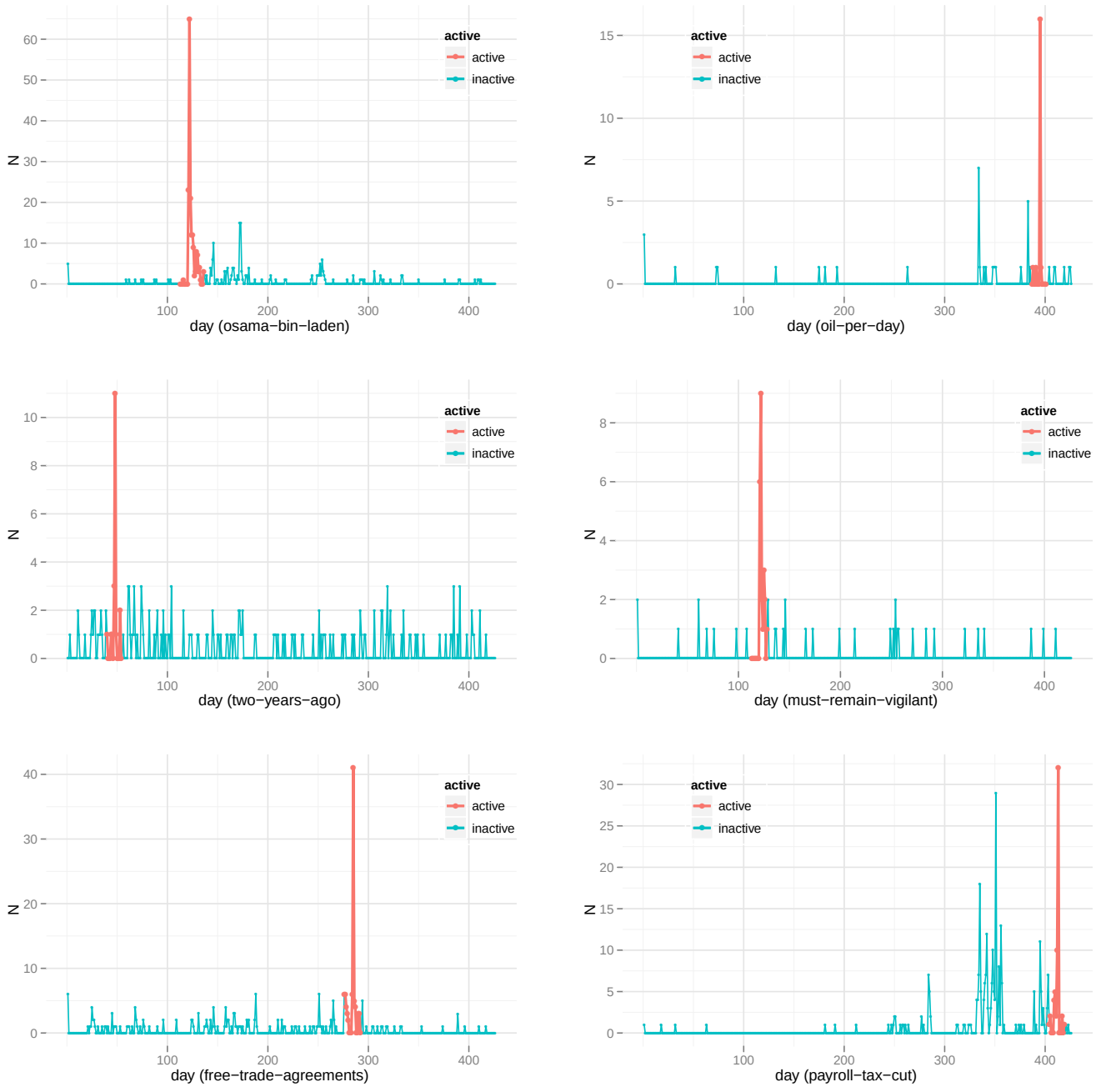

Figure 4. Examples of trigram infection instances. Note that a trigram may be associated with multiple non-overlapping bursts (e.g. "payroll-tax-cut" around day 350 and day 410), and each of the non-overlapping bursts is considered as a different infection instance.

by Republicans to refer to the failures of the Obama administration during the first half of the President's term. The burst is clearly distinct from the regular usage of the phrase, and also suggests a modest, temporary increase in the use of the phrase in subsequent weeks. Eventually the use of this frame appears to die down to pre-burst levels.

Figure 4(d) shows the burst in the use of the frame "must remain vigilant." The use of this frame has a absolute mean increase .82. Prior to the burst period, covering 120 days, the phrase was used 9 times by Senators in their public statements. It was then used 24 times during the burst period. The rate of use seems to return to pre-burst levels shortly thereafter, however.

Finally, Fig. 4(f) shows the use of the phrase "payroll tax cut." This phrase was jointly agreed upon by Democratic senators in their joint press release. The burst in usage represents their jointly authored letter

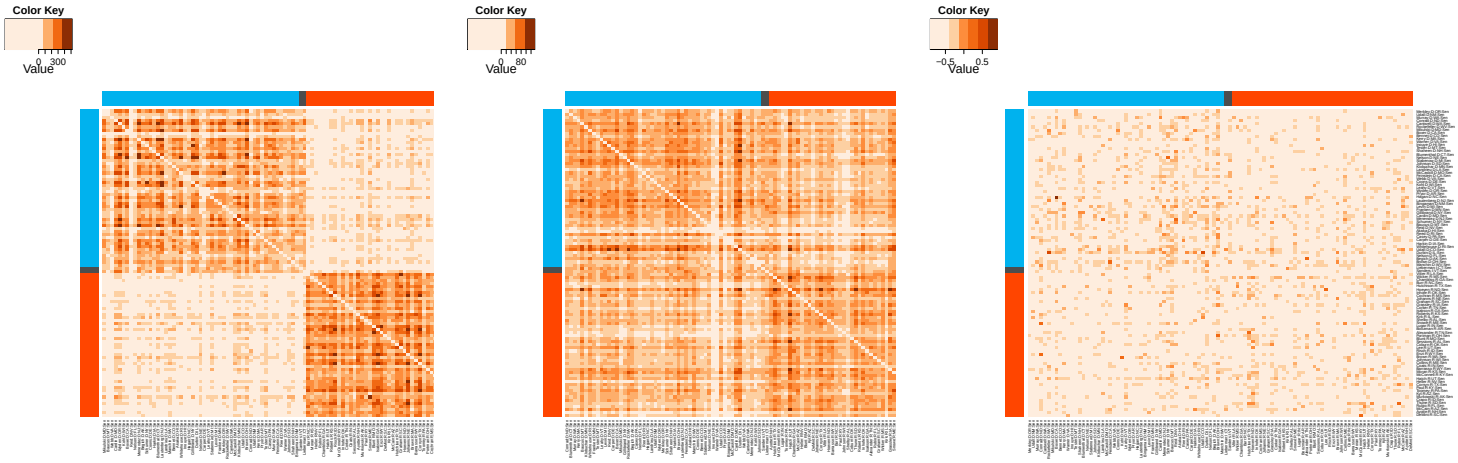

Figure 5. The matrices of the three different social networks extracted from the dataset: (a) Joint Authoring, (b) Infection Sharing and (c) Infection Following networks. Rows and columns in the matrices are reshuffled according to party and connections. The colors around the rows and columns indicate the party affiliation (Republican: red; Democrat: blue).

to the Senate. This burst showed a 2.64 absolute mean increase usage. This appears to be the third time that this phrase showed a burst. Interestingly, the prior bursts did not contain a large, joint authored press release. This suggests that the phrase had established itself as a legitimate frame prior to being one that Senators would deploy in a jointly authored document.

Fig. 7 shows the examples of actual texts appearing in Senators' press release after Bin Laden's death. The phrase "must remain vigilant" was used to implore the continued commitment to anti-terror efforts. Previously, this phrase had rarely been used in the public statements of Senators. We detected that 6 senators used the phrase on May 1, 2011[9], followed by 9 senators the next day[10], and then another 8 over the next several days[11]. The application of this phrase did not end there, however. On the ninth day after Bin Laden's death, one Senator deployed the phrase "must remain vigilant" in a new context. Senator John Boozman of Arkansas cautioned that it was important to intensely prepare for natural disasters, stating "we all must remain vigilant to protecting ourselves and our loved ones this storm season" (Boozman, May, 9, 2011). This example provides evidence that repetition of the phrase around one event legitimated its use for other contexts.

5.2. Analysis on Extracted Networks

As shown in Fig. 5, three types of social networks are extracted: Joint Authoring, Infection Sharing and Infection Following networks. Fig. 6 shows the Infection Following Network of Senators. Below we present the analysis results for comparing the structures of these extracted networks with one another and with additional information gathered for the Senate.

Comparing the Inferred Networks. Both the individual links and the centrality scores were compared for the three inferred networks. Centrality was computed using both a raw sum of the column scores for each senator as well as the page rank score for each senator. Both Pearson correlations and Spearman's rank correlations for these scores were extremely high (within the same network, $> .99$ for each network). Thus, only the page rank score is reported. Correlations between network links were computed using the Quadratic Assignment Procedure (QAP) [72].

Infection Following Networks vs. Infection Sharing Network. These networks are built from the same data but are calculated differently. In the infection following network, infections flow in a time ordered manner. Individuals that use a bursting trigram first receive in-bound links from individuals that use the same bursting trigram at a later point in time. The weight of this link is determined by a decay function and weakens as time passes between the first individual's usage and the second's. The weight of this link is also determined by the number of other individuals that the second individual is "following." The first individual receives a stronger link from the second if he/she is the only senator that used this trigram prior to the second individual's doing so. If the second individual's usage follows a large group of senators, each senator receives a weak link only. The infection sharing network does not use any weighting. The fact that two senators each used a bursting trigram within the burst window is sufficient to give them a symmetric tie of value (count) 1. Thus the infection sharing network tends to capture shared persuasive interests – the fact that two senators

[9] For example, see http://www.cardin.senate.gov/newsroom/statements_and_speeches/statement-by-us-senator-ben-cardin-on-death-of-osama-bin-laden

[10] For example, see http://conrad.senate.gov/pressroom/record.cfm?id=332649

[11] For example, see http://www.crapo.senate.gov/media/newsreleases/release_full.cfm?id=332701

chose to use the same bursting trigram – but discards information regarding innovation and diffusion, which are inherently temporal [54].

Comparing these networks shows a strong association between the two but also some differences. The centrality scores of the two networks are highly correlated ($r = .84, p < .001, df = 98$). This suggests that individuals that tend to be early and alone in bursts and are thus central in the following network are more consistent participants in bursts, whereas followers tend to be a more heterogeneous group. Thus, when followers are given equal weight to innovators (as in the infection sharing network), centrality ranking does not change substantially. Table 1 displays the top 10 senators in terms of page rank centrality for each of the three networks.

The two networks are not identical, however. The link to link correlation as calculated by the QAP procedure shows $r = .325$ ($p < .001$, using 1000 permutations). This suggests that removing timing and weights leads to substantial changes in the network.

Infection Sharing Network vs. Joint Authorship Network. These networks include several instances of overlapping data but also distinct data points. When a joint authored document uses a trigram that participates in a burst, both networks will include this document and its authors will receive links to one another in both networks.

The infection sharing network also includes links between senators that did not co-author documents, while the joint authorship network also includes documents that did not use any bursting phrases and thus did not qualify for the infection networks.

The correlation between the network links is significant ($r = .38$, $p < .001$, using 1000 permutations). This appears sensible given that both networks share several documents in common and assign links on these shared documents in the same way. However, comparing these networks shows only a weak association in network centrality ($r = .17, p = .08, df = 98$). This suggests that individuals that obtain additional links in the symmetric network, by virtue of participating in solo-authored public statements using bursting trigrams, are not particularly likely to participate in other, joint-authored statements with non-bursting phrases. This could indicate that some individuals are simply more likely to communicate through solo-authored statements rather than through joint-authored statements. This may also be due to the fact that the joint authorship network shows a high degree of connectedness amongst many authors, leading many senators to have high centrality scores that are not meaningfully different from one another.

Infection Following Network vs. Joint Authorship Network. The correlation between the following infection network and the joint authorship network shows a different pattern. The centrality scores show almost no correlation ($r = .001, p = .99, df = 98$). This suggests that those likely to be followed are not those that are likely to co-author with others. This may be explained by the fact that some senators choose to communicate as solo authors whereas others choose to participate in group-authored documents.

Yet the networks show a significant but modest correlation ($r = .12, p < .001$, based on 1000 permutations). The network correlation is interesting because these two networks do not derive links from the same documents in the same way. For a joint-authored document that participates in a semantic burst, each author will receive a link to each other author for that document. They will not receive links to one another in the infection following network, however, because the co-authored usage of the bursting term is simultaneous in our data. The correlation between these networks suggests that individuals that participate in co-authored documents also obtain links from one another in other or subsequent cases. It is also possible that this correlation is simply due to common covariance with a third structure. This possibility is explored in more detail in the next section.

Comparison to Covariate Networks. This section reports results of QAP correlation analyses between the inferred networks and the covariate networks. Table 2 below shows the correlations and the significance levels. The results indicate that there is a substantial correlation between party membership and the inferred networks. There are also modest but significant correlations between the inferred networks and both committee membership and geographic adjacency of home states.

The joint authorship network shows a very strong link to party membership, but no association with shared committee membership. This suggests that collaboration is distinct from pure exposure or working together on common issues.

The results of the network correlations suggests that the bursts and burst participants detected by the method may reveal some latent coordinative and cooperative relationships that are distinct from party membership and observable collaboration. One question is whether the significant association between the infection following network and the joint authorship network remains significant when controlling for known factors that suggest senators have common interests – shard party and similar geographic origins (as occurs when they represent adjacent states). Another question is whether the significant relationships between committee membership and adjacent home states to infection following and infection sharing remain after controlling for party membership.

To test answer these questions a series of MRQAP regressions were run [73]. This technique calculates the partial correlations between a predicted network

and predictor networks as in a regression. To answer the first question, the joint authorship network was regressed on the infection following network, the party affiliation network and the adjacent states network. The coefficient for infection following remained significant ($p < .01$), though the additional variance in the joint authorship network was reduced to .001. This suggests that a substantial portion of the shared variance between these networks was due to covaration with factors already known to indicate shared interests – party affiliation and state adjacency – however a small but significant portion remains suggesting latent affinities between particular senators.

To answer the second question, the infection following network was regressed on the party affiliation network, the committee membership network, and the geographic adjacency network. All three networks continue to show a significant correlation with the infection following network (shared party affiliation $p < .001$, shared committee membership $p < .001$, geographic adjacency $p < .001$). These results suggest that the infection following network captures some informal communication pattern that is not explained strictly by party affiliation.

Attribute Correlations to Network Centrality. The following analyses report first order correlations between inferred network centrality scores and the individual attribute scores. For each attribute a primary correlation is calculated using all senators and then a subset of analysis are conducted for Democrats linking to Democrats, Republicans linking to Republicans, Democrats linking to Republicans and Republicans linking to Democrats. These subsets of the overall correlation allow for validation that there are consistent underlying processes.

Year Entered Senate. Table 3 displays the correlation in network centrality for each network with the year the senator entered the senate. A positive correlation indicates that a more junior senator is more central.

The results show a significant but distinct relationship for both the infection following network and the joint authorship network. More senior senators are more likely to be followed by others, meaning that a bursting trigram used by a senior senator is likely to be used by other senators on subsequent days. These senior senators are less central in the joint authorship process, however. They may jointly author fewer press releases or they may consistently select only small number of partners.

Examining the intra-party and cross-party dynamics shows similar results. For three out of the four subsets, a similar pattern is observed. For each of Rep-Rep, Dem-Rep, and Rep-Dem links the correlation between Year Entered and infection following centrality is negative and between Year Enter and joint authorship centrality is positive, with all of the coefficients > .10 in absolute magnitude and statistical significance for 2 measures (Democrats following Republicans, $r = -.41, p < .001, df = 43$; and Republicans jointly authoring with Democrats, $r = .52, p < .001, df = 49$).

These results suggest that seniority gives an individual communicative authority. However, amongst Democrats relations to other Democrats, the pattern is different. Following of senior senators is no longer significant ($r = -.07, p = .63, df = 49$), while joint authorship also appears to favor more senior senators ($r = -356, p < .05, df = 49$).

Leadership. Table 4 displays the correlation in network centrality for each network with the a dummy variable for whether the senator holds a leadership position.

Despite the fact that only 6 senators are coded with this dummy, the results show substantial correlations. Consistent with the findings regarding seniority, senators in leadership positions appear to be less central in the joint authoring of press releases. The relationship between following and leadership is less pronounced and is not statistically significant. The reasons become somewhat clear when examining the subsets based on party identification.

Table 5 shows each of the four subset networks and the correlation scores. These results suggest that Republicans show more consistent behavior with regard to leadership. They are significant in their tendency to follow Republican leadership and significant in their tendency not to jointly author documents with Democratic leadership.

Term Ending Year. Table 6 displays the correlation in network centrality for each network with the year the senator will be up for re-election. A positive correlation indicates that a senator whose re-election campaign is further in the future is more central. This measure thus captures the extent to which senators' public statements may be motivated by a more imminent re-election campaign.

This table shows no significant relationships. This may be due to the fact that the majority of the statements included in the analysis were taken from 2011 during which the nearest re-election campaign was still a year away. Analysis of the subsets shows no significant correlations or consistent patterns across subsets.

Ideology. Table 7 displays the correlation in network centrality for each network with the senator's ideology score as calculated by the DW-Nominate algorithm. A positive correlation indicates that a senator who is more conservative is more central, a negative correlation indicates that a senator who is more liberal is more central.

As might be expected, there is no consistent relationship between ideology and being followed or

Table 2. QAP Correlations with Covariate Networks

	Follow	Share	Joint
Shared Party	.12 ***	.31 ***	.71 ***
Committee Membership	.04 ***	.05 *	.02
Adjacent States	.03 **	.05 ***	.11 ***

$*p < .05, **p < .01, ***p < .001$; permutations = 1000

Table 3. Correlation between Network Centrality and Year Entered Senate for all senators

	Follow	Share	Joint
Year Entered	-0.218 *	-0.109	.368 ***

$*p < .05, **p < .01, ***p < .001, df = 98$

Table 6. Correlation between Network Centrality and Term Ending year for all senators

	Follow	Share	Joint
Term End	-0.108	-0.091	0.087

$*p < .05, **p < .01, ***p < .001, df = 98$

Table 7. Correlation between Network Centrality and DW-Nominate for all senators

	Follow	Share	Joint
DW-Nominate	-0.178	-0.018	0.114

$*p < .05, **p < .01, ***p < .001, df = 98$

co-authored with. This is likely due to the partisan structure of the Senate.

Table 8 shows the effect of ideology within the subsets. The correlation between ideology and the joint authorship network appears to be quite strong. As would be expected, Republicans tend to jointly author public statements with conservative Democrats and Democrats tend to jointly author public statements with liberal Republicans. It also appears that joint authorship centers on individuals in the political extremes of the parties, with Democrats favoring liberal Democrats and Republicans favoring conservative Republicans. The relationships for infection following are more difficult to parse. The one significant relationship is for Democrats following more liberal Democrats.

6. Discussion

Review of Findings We described four basic mechanisms which might lead politicians to show rapid semantic convergence. Our method of examining bursts in trigram usage suggests that each of these mechanisms appears to operate on the communication behavior of U.S. Senators at least part of the time. While we have not examined each individual detected burst to determine which mechanism most likely gave rise to it, the examples we have examined suggest that these theoretical categories are a good starting point for further investigation.

Outside of the convergence due to shared basic categories and names, the other processes rely on some form of social coordination. We capture these social coordination mechanisms broadly with three distinct networks: the infection following network, the infection sharing network, and the jointly authored network. The first two are built from the detection of semantic bursts, while the jointly authored network is constructed from the detection of identical documents shared by different authors.

Analyses of these networks suggests that identifying rapid semantic convergence reveals a meaningful social structure. The infection following network showed an independent relationship to the shared committee membership and geographic adjacency network even when party affiliation was controlled for. This suggests that the relationship has some basis in shared interests that may be issue specific or common exposure to frames or arguments. Exposure may be through face-to-face interaction via committee work or joint authorship or through media that are common to several senators home states. The fact that the infection following network showed a small but significant correlation to the joint authorship network suggests that face-to-face interactions may play a role.

The network centrality scores for the infection following network and the joint authored network also revealed some interesting patterns. Results suggested that more senior senators and senators in leadership positions were more likely to participate early in semantic bursts and be subsequently followed in the usage of bursting terms by others. By contrast, these individuals were less likely to jointly author with other individuals.

One explanation for this phenomenon is that these more senior senators do not need to establish ties to other senators in order to have a credible or authoritative voice. In a sense, it may require a co-authored letter or press release by several junior senators to achieve the same impact as a senator in a leadership position might achieve with a solo-authored document. Research in inter-organizational partnerships suggests that this ability to operate free of the constraints of others is an indicator of an organization's power [74, 75].

Another explanation for this phenomenon is that more junior senators and senators without leadership positions jointly author many statements that do not qualify for bursts. If this were the case, they could acquire central positions in the joint network without our method finding that they have followers. While this phenomenon does not account for the significance of

Table 1. Top Ten by Page Rank (infector)

Infections Followed (by others)		Infections Shared (with others)		Joint Authored Documents	
Name: Party: State	PR Score	Name: Party: State	PR Score	Name: Party: State	PR Score
Brown:D:OH	0.0211	Durbin:D:IL	0.0160	Wyden:D:OR	0.0156
Hatch:R:UT	0.0193	Hatch:R:UT	0.0160	Brown:D:OH	0.0149
Durbin:D:IL	0.0191	Brown:D:OH	0.0145	Isakson:R:GA	0.0143
Reid:D:NV	0.0180	Hutchison:R:TX	0.0140	Johanns:R:NE	0.0140
Baucus:D:MT	0.0179	Cardin:D:MD	0.0140	Klobuchar:D:MN	0.0136
Cardin:D:MD	0.0177	Rockefeller:D:WV	0.0135	Johnson:R:WI	0.0135
Leahy:D:VT	0.0175	Baucus:D:MT	0.0134	Ayotte:R:NH	0.0135
Schumer:D:NY	0.0163	Snowe:R:ME	0.0133	Cornyn:R:TX	0.0134
Boxer:D:CA	0.0163	Whitehouse:D:RI	0.0132	Cardin:D:MD	0.0133
Rockefeller:D:WV	0.0159	Boxer:D:CA	0.0132	Menendez:D:NJ	0.0131

Table 4. Correlation between Network Centrality and Leadership position for all senators

	Follow	Share	Joint
Leader	0.185	0.133	-0.234

$*p < .05, **p < .01, ***p < .001, df = 98$

Table 5. Correlation between Network Centrality and Leadership position for party subsets

Subset	Variable	Follow	Share	Joint	df
Democrats Following Democrats	Leader	0.186	0.098	-0.21	49
Republicans Following Republicans	Leader	.31 *	0.22	-0.142	45
Democrats Following Republicans	Leader	0.083	0.093	-0.21	43
Republicans Following Democrats	Leader	0.084	0.124	-0.36 **	49

Table 8. Correlation between Network Centrality and DW-Nominate for party subsets

Subset	Follow	Share	Joint	df
Democrats Following Democrats	-0.386 **	-0.41 **	-0.542 ***	48
Republicans Following Republicans	0.048	0.073	.336 *	32
Democrats Following Republicans	-0.327	-.340 *	-.486 **	48
Republicans Following Democrats	-0.199	-0.238	.278 *	32

$*p < .05, **p < .01, ***p < .001$

more senior members as leaders, it suggests a partial explanation for the observed discrepancy.

Our results also suggest some interesting similarities and differences between the parties. As would be expected, members of both parties are more likely to jointly author statements with members of the opposing party that are closer to them ideologically. Members of both parties also seem to jointly author statements within the party with members that represent the ideological extremes. This may be additional evidence of the recent polarization in American politics [44, 76].

Following relationships are not as clear cut, however, as most results within and across parties were not statistically significant. One exception is the behavior of Republicans with respect to party leadership. Republicans are significantly more likely to follow their own leaders, and significantly less likely to jointly author statements with Democratic leaders. This may be because Republicans, in general, are more disciplined or because, as the minority party in the 112th Senate, they find it necessary to be more strategic in their communication.

There were no significant results for the extent to which a Senator's re-election was imminent. This may be due to the fact that the majority of the data reflect a period more than a year before the next election.

Limitations. This study develops a new method based on a theoretical understanding of semantic convergence. As such the methodology has yet to be refined and poses some limitations. First, as our method was exploratory, it has not been refined to distinguish between the theoretical mechanisms which lead to semantic convergence. The method identified numerous examples which we manually analyzed to determine how they fit into the theoretical framework. Further testing is required to see whether this method is sufficiently sensitive to convergence due to each

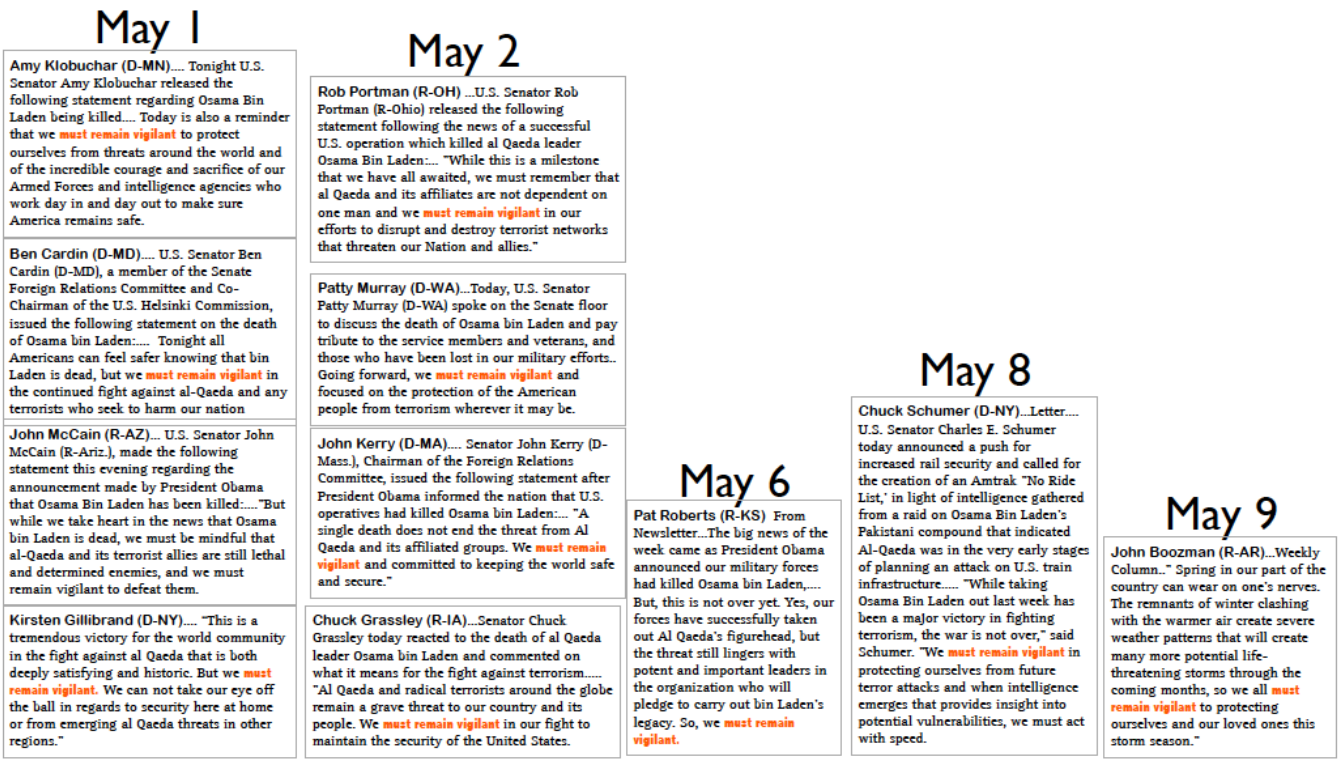

Figure 7. The trigram "must remain vigilant" appeared frequently in the press release over the week after Bin Laden's death.

mechanism, and at this point no claim regarding the accuracy of the tool can be made.

Another limitation in this approach is the reliance on temporal measurement at the daily level. News cycles are often very short and it may be that many important processes are taking place within a single day. At the same time, many important communication decisions may be made over long periods of time and day-to-day distinctions between when senators use particular phases may be theoretically meaningless. That is, the individuals that use phrases first may simply have more efficient press secretaries. The high correlation between the centrality scores for the infection following and the infection sharing networks suggest that such a process is unlikely to be distorting the results in this study, however it remains a limitation of a method that relies on data that is time stamped in a manner that may not be theoretically appropriate.

Thirdly, the conclusions drawn from the analyses of these networks must remain tentative as the data were collected for only a small number of individuals (100 Senators) over a particular period of time. The modest sample size, particularly for variables such as leadership and ideology, particularly when applied within parties, substantially limits the statistical power of the analyses. Furthermore, the fact that these data reflect a particular point in time may have introduced noise from particular events or circumstances in that time period.

Suggestions for Further Research. Based on these initial findings we suggest the following avenues for further research. First, a variety of techniques may be useful in helping to distinguish the theoretical causes of semantic convergence. Potential techniques include the use of sentiment indicators to distinguish framed or persuasive language from categorical or naming language. It may also be possible to identify categories and names by using a dictionary of proper names or words that indicate political entities, such as "act" "office" and "department." Other techniques might involve inspecting the distributions before, during and after bursts. There may be particular signatures associated with the manner in which frames gain adherents as opposed to entities or evidence.

Adding additional data from the House of Representatives and other session of congress is also likely to improve the analysis. Additional observations will add statistical power as well as offer the opportunity to test for longitudinal dynamics, including the influence of elections and the development of following and collaboration relationships over time.

The interesting results regarding the joint press-release network also suggest these joint press releases could be analyzed in their own right. The individual predictors and emergent structure for this behavior may

complement other analyses of political collaboration, such as those done on co-sponsorship networks.

Analyses of the role of surrounding media as both an initiator and follower of bursts in politicians' statements will also be of interest. Substantial research has examined opinion leadership and agenda setting by treating media coverage as a dependent variable [7, 40, 76, 77]. Little research has examined the ways in which politicians may themselves be influenced by media coverage. While politicians with focused agendas and a specific set of targeted constituents may not adjust their positions in response to media trends or news cycles, they may still rely on the media to fortify them with arguments, evidence and rhetorical innovations which they can use to advocate for their positions. The method described in this article offers a means of detecting such cases should they occur.

7. Conclusion

In this article we presented a method of detecting strategic communication coordination amongst politicians through the detection of semantic convergence in the form of bursts in the use of particular phrases. A variety of analyses suggest that the technique is able to identify otherwise difficult to observe relationships and alliances amongst political actors, as well as structural patterns that suggest the dynamics of cooperative behavior among the community members.

Acknowledgements. We gratefully acknowledge the support of the Lazer Lab at Northeastern University, supported in part by MURI grant #504026, DTRA grant #509475, and ARO #504033. We thank Sasha Goodman for his early assistance with data collection. We also thank all our colleagues who provided insight and comments that greatly assisted the research.

References

[1] Engestrom, Y. (1992) *Interactive Expertise: Studies in Distributed Working Intelligence. Research Bulletin 83.* (ERIC). URL http://eric.ed.gov/?id=ED349956.

[2] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A., Brewer, D., Christakis, N. *et al.* (2009) Computational social science. *Science* **323**(5915): 721.

[3] Roth, C. and Cointet, J. (2010) Social and semantic coevolution in knowledge networks. *Social Networks* **32**: 16–29.

[4] Strang, D. and Meyer, J. (1993) Institutional conditions for diffusion. *Theory and Society* **22**: 487–511.

[5] Chong, D. and Druckman, J.N. (2007) Framing theory. *Annual Review of Political Science* **10**(1): 103–126.

[6] Hopmann, D.N., Vliegenthart, R., De Vreese, C. and Albæk, E. (2010) Effects of election news coverage: How visibility and tone influence party choice. *Political Communication* **27**(4): 389–405. Doi: 10.1080/10584609.2010.516798.

[7] Dunn, S.W. (2009) Candidate and media agenda setting in the 2005 virginia gubernatorial election. *Journal of Communication* **59**(3): 635–652.

[8] Benford, R.D. and Snow, D.A. (2000) Framing processes and social movements: An overview and assessment. *Annual Review of Sociology* **26**: 611–639.

[9] Brummans, B., Putnam, L., Gray, B., Hanke, R., Lewicki, R. and Wiethoff, C. (2008) Making sense of intractable multiparty conflict: A study of framing in four environmental disputes. *Communication Monographs* **75**(1): 25–51.

[10] Bimber, B. (2003) *Information and American Democracy: Technology in the Evolution of Political Power* (New York: Cambridge University Press).

[11] Slothuus, R. (2010) When can political parties lead public opinion? evidence from a natural experiment. *Political Communication* **27**(2): 158–177. Doi: 10.1080/10584601003709381.

[12] Grimmer, J. (2009) A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis* **18**(1): 1–35.

[13] Gandy, O. (1982) *Beyond Agenda Setting: Information Subsidies and Public Policy* (Norwood, NJ: Ablex.).

[14] Beckwith, R., Fellbaum, C., Gross, D. and Miller, G. (1991) Wordnet: A lexical database organized on psycholinguistic principles. *Lexical acquisition: Exploiting on-line resources to build a lexicon* : 211–232.

[15] Vigliocco, G. and Vinson, D. (2007) Semantic representation. *The Oxford handbook of psycholinguistics* : 195.

[16] Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshman, R. (1990) Indexing by latent semantic analysis. *Journal of the American society for information science* **41**(6): 391–407.

[17] Hofmann, T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* **42**(1): 177–196.

[18] Blei, D., Ng, A. and Jordan, M. (2003) Latent dirichlet allocation. *The Journal of Machine Learning Research* **3**: 993–1022.

[19] Salton, G. and Buckley, C. (1988) Term-weighting approaches in automatic text retrieval. *Information processing & management* **24**(5): 513–523.

[20] Zhu, Y. and Shasha, D. (2003) Efficient elastic burst detection in data streams. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM): 336–345.

[21] Vlachos, M., Meek, C., Vagena, Z. and Gunopulos, D. (2004) Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data* (ACM): 131–142.

[22] Swan, R. and Allan, J. (1999) Extracting significant time varying features from text. In *Proceedings of the eighth international conference on Information and knowledge management* (ACM): 38–45.

[23] Basseville, M., Nikiforov, I. *et al.* (1993) *Detection of abrupt changes: theory and application*, **15** (Prentice Hall Englewood Cliffs).

[24] Kleinberg, J. (2003) Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery* **7**(4): 373–397.

[25] Mane, K. and Börner, K. (2004) Mapping topics and topic bursts in pnas. *Proceedings of the National Academy of Sciences of the United States of America* **101**(Suppl 1): 5287.

[26] Carley, K.M. (1997) Extracting team mental models through textual analysis. *Journal of Organizational Behavior* **18**(s 1): 533–558.

[27] Danowski, J.A. (2011) *Counterterrorism Mining for Individuals Semantically-Similar to Watchlist Members* (Springer).

[28] Searle, J. (1969) *Speech Acts: An introduction to the philosophy of language* (Cambridge: Cambridge).

[29] Murphy, G.L. (2002) *The Big Book of Concepts* (Cambridge, MA: MIT Press).

[30] Rosch, E., Mervis, C., Gray, W., Johnson, D. and Boyes-Braem, P. (1976) Basic objects in natural categories. *Cognitive Psychology* **8**: 382–439.

[31] Corter, J. and Gluck, M. (1992) Explaining basic categories: feature predictability and information. *Psychological Bulletin* **111**(2): 291–303.

[32] Rosch, E. (1978) *Principles of categorization* (Hillsdale, NJ: Lawrence Erlbaum), 27–48.

[33] Steyvers, M. and Tenenbaum, J.B. (2005) The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science* **29**(1): 41–78. Times Cited: 107.

[34] Lynch, E., Coley, J. and Medin, D. (2000) Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory & Cognition* **28**(1): 41–50.

[35] Medin, D., Lynch, E., Coley, J. and Atran, S. (1997) Categorization and reasoning among tree experts: Do all roads lead to rome? *Cognitive Psychology* **32**: 49–96.

[36] Proffitt, J., Coley, J. and Medin, D. (2000) Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26**: 811–828.

[37] Margolin, D.B. and Monge, P. (2013) Conceptual retention in epistemic communities. In Moy, P. [ed.] *Communication and community* (New York: Hampton Press), 1–22.

[38] Hänggli, R. and Kriesi, H. (2010) Political framing strategies and their impact on media framing in a swiss direct-democratic campaign. *Political Communication* **27**(2): 141–157. Doi: 10.1080/10584600903501484.

[39] Berger, J. and Heath, C. (2005) Idea habitats: How the prevalence of environmental cues influences the success of ideas. *Cognitive Science* **29**: 195–221.

[40] Tedesco, J.C. (2005) Intercandidate agenda setting in the 2004 democratic presidential primary. *American Behavioral Scientist* **49**(1): 92–113.

[41] Fowler, J.H. (2006) Legislative cosponsorship networks in the us house and senate. *Social Networks* **28**(4): 454–465.

[42] Petty, R. and Cacioppo, J. (1986) *Communication and persuasion: Central and peripheral routes to attitude change* (New York: Springer- Verlag).

[43] Vraga, E.K., Edgerly, S., Wang, B.M. and Shah, D.V. (2011) Who taught me that? repurposed news, blog structure, and source identification. *Journal of Communication* **61**(5): 795–815.

[44] Baum, M.A. and Groeling, T. (2008) New media and the polarization of american political discourse. *Political Communication* **25**(4): 345–365.

[45] Medvetz, T. (2006) The strength of weekly ties: Relations of material and symbolic exchange in the conservative movement. *Politics & Society* **34**(3): 343–368.

[46] Entman, R. (1993) Framing: toward clarification of a fractured paradigm. *Journal of Communication* **43**: 51–58.

[47] Murphy, G.L. and Medin, D. (1985) The role of theories in conceptual coherence. *Psychological Review* **92**(3): 289–316.

[48] Lakoff, G. (1987) *Women, fire, and dangerous things: What categories reveal about the mind* (University of Chicago press).

[49] Weber, K., Heinze, K. and DeSoucey, M. (2008) Forage for thought: Mobilizing codes in the movement for grass-fed meat and dairy products. *Administrative science quarterly* **53**(3): 529.

[50] Seibold, D. and Meyers, R. (2007) Group argument: A structuraion perspective and research program. *Small Group Research* **38**: 312–336.

[51] March, J. and Simon, H. (1993) *Organizations* (Cambridge, MA: Blackwell), 2nd ed.

[52] Thagard, P. (1997) *Coherent and creative conceptual combinations.* (Washington, D.C.: American Psychological Association).

[53] Wisniewski, E. (1996) Construal and similarity in conceptual combination. *Journal of Memory and Language* **35**: 434–453.

[54] Rogers, E. (2003) *Diffusion of Innovations* (New York: The Free Press), 5th ed.

[55] Sorenson, O., Rivkin, J.W. and Fleming, L. (2006) Complexity, networks and knowledge flow. *Research Policy* **35**(7): 994–1017. Times Cited: 36.

[56] Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chavez, C. and Wojick, D. (2008) Population modeling of the emergence and development of scientific fields. *Scientometrics* **75**: 495–518.

[57] Lambiotte, R. and Panzarasa, P. (2009 in press) Communities, knowledge creation, and information diffusion. *Journal of Informetrics* .

[58] Monge, P. and Poole, M. (2008) The evolution of organizational communication. *Journal of Communication* **58**: 679–692.

[59] Phillips, N., Lawrence, T. and Hardy, C. (2004) Discourse and institutions. *Academy of Management Review* **29**: 635–652.

[60] Ghaziani, A. and Ventresca, M. (2005) Keywords and cultural change: Frame analysis of business model in public talk, 1975-2000. *Sociological Forum* **20**(4): 523–559.

[61] Clark, H. and Wilkes-Gibbs, D. (1986) Referring as a collaborative process. *Cognition* **22**: 1–39.

[62] Nelson, R. and Winter, S. (1983) *An Evolutionary Theory of Economic Change* (Cambridge, MA: Belknap).

[63] Weick, K. (1979) *The Social Psychology of Organizing* (Reading, MA: Addison-Wesley).

[64] Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J., Hoiberg, D. *et al.* (2011) Quantitative analysis of culture using millions of digitized books. *science* **331**(6014): 176–182.

[65] Liu, C. and Srivastava, S. (in press) Pulling closer and moving apart: Interaction, identity, and influence in the u.s. senate, 1973-2009. *American Sociological Revie* .

[66] Krackhardt, D. (1994) *Constraints on the Interactive Organization as an Ideal Type* (Thousand Oaks, CA: Sage Publications), 211–222.

[67] Huang, M., Contractor, N., Huang, Y., Margolin, D., Ognyanova, K. and Shen, C. (2010), The effects of diversity and repeat collaboration on performance in distributed nanoscientist teams.

[68] Gentzkow, M. and Shapiro, J.M. (2010) What drives media slant? evidence from us daily newspapers. *Econometrica* **78**(1): 35–71.

[69] Bennett, W. and Iyengar, S. (2008) A new era of minimal effects? the changing foundations of political communication. *Journal of Communication* **58**(4): 707–731.

[70] Althaus, S.L., Cizmar, A.M. and Gimpel, J.G. (2009) Media supply, audience demand, and the geography of news consumption in the united states. *Political Communication* **26**(3): 249–277. Doi: 10.1080/10584600903053361.

[71] Carroll, R., Lewis, J.B., Lo, J., Poole, K.T. and Rosenthal, H. (2011), Dw-nominate scores with bootstraped standard errors.

[72] Krackhardt, D. (1987) Qap partialling as a test of spuriousness. *Social Networks* **9**(2): 171–186. Times Cited: 82.

[73] Dekker, D., Krackhardt, D. and Snijders, T. (2007) Sensitivity of mrqap tests to collinearity and autocorrelation conditions. *Psychometrika* **72**(4): 563–581.

[74] Pfeffer, J. and Salancik, G. (1978) *The external control of organizations* (New York: Harper Row).

[75] Powell, W.W., White, D., Koput, K. and Owen-Smith, J. (2005) Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American Journal of Sociology* **111**(5): 1463–1568.

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

[77] Meraz, S. (2009) Is there an elite hold? traditional media to social media agenda setting influence in blog networks. *Journal of Computer-Mediated Communication* **14**(3): 682–707.