From Words to Networks and Back:

Digital Text, Computational Social Science,

and the Case of Presidential Inaugural Addresses*

Ryan Light

University of Oregon

Word Count: 9,687

This is a postprint. The version of record is available online at:
http://journals.sagepub.com/doi/full/10.1177/2329496514524543

*The author thanks James Moody, Jill Ann Harrison, jimi adams, Pamela Paxton and the editors and reviewers for their helpful comments on this paper. Please direct all correspondence to Ryan Light, Department of Sociology, 1291 University of Oregon, Eugene, OR 97403 light@uoregon.edu.

From Words to Networks and Back: Digital Text, Computational Social Science, and the Case of Presidential Inaugural Addresses

Abstract: Digital text has revolutionized how we consume and produce information and also provides seemingly limitless sources of data from twitter feeds to online historical archives. Such new data challenge traditional boundaries between quantitative and qualitative research and exciting horizons have emerged. New analytic approaches are warranted, however, given the typically unstructured, respondent-generated format of such data. In this article, I examine how sociologists have handled text data prior to digitization. Building on recent advancements in computational linguistics and computational social science, I then offer a network-based model and approach for analyzing text similarities and locating emergent, general themes in a relatively systematic way. I provide a case in point by analyzing United States' presidential inaugural addresses. This analysis illustrates how sociologists can take advantage of both the breadth of new digital sources of data and the richness that such qualitative material provides. Indeed, the digitization of texts represents a possible and stimulating sea change in how we tell socio-cultural and historical stories. The greatest potential in these regards rests at the nexus of new computational methods and in-depth, qualitative strategies.

Keywords: Cultural Sociology, Comparative/Historical Methods, Text Analysis, Computational Social Science

The death of the novel. The death of the magazine. The death of the newspaper.[1] The mortality of text-based cultural objects has long seemed right around the corner. While paper media may be suffering a decline amid years-old signs of extinction, text-based media itself has never been so popular. This is somewhat counterintuitive given the promise of new virtual forms of communication, but truth be told we are not quite up to the challenge yet. Internet communication has only brought more text information into our lives. We write letters in the form of email, we pass notes in the form of status updates on social networking sites, we post and read diaries on blogs. Digital libraries are growing exponentially, bringing books directly into our homes without the threat of overdue notices (Grafton 2007). We are, more or less, digital Victorians.

How well have the social sciences, and sociology in particular, adapted to this new-found textuality? We see certain growth in hybrid fields, like computational linguistics and the information sciences. There has been cross-fertilization of these fields into English through linguistics (e.g. Collins et al. 2004); political science through political psychology and discourse (e.g. Simonton 1988; Young 1996; see Cousins and McIntosh 2005 for a discussion); and sociology through social linguistics and cultural analyses (e.g. Bearman, Faris, and Moody 1999; Smith 2007; see Mohr 1998 and Mohr and Bogdanov 2013 for a discussion). We also see some interest in text within sociological areas as diverse as political sociology and social movements through the use of newspaper data (see Earl et al. 2004 for a review) and in the sociology of science (e.g. Moody and Light 2006), among numerous others. However, the social sciences are behind other disciplines in exploring digitized text, prompting a call (e.g., Lazer et al. 2009) for more concerted efforts to build an interdisciplinary computational social science.

[1] There are numerous examples of critics and scholars predicting the decline of print media. For a few examples, see Fitzpatrick (2006) for a description of various claims about the novel's decline and *The Economist* (2006) on the obsolescence of print news media.

How might sociologists play a role in the nascent computational social science specifically related to the analysis of text data? In this article, and following a brief overview of how sociologists have handled text data prior to digitization, I describe several recent and promising strategies. Building on them, I offer a network-based model for analyzing text similarities and emergent themes within text-based corpora. I illustrate these techniques in an analysis of United States' presidential inaugural addresses. My example—an example that shows how texts in a speech-to-speech network are organized in two distinct clusters by time, that reveals main thematic clusters, and that denotes shifts over time such as a growing concern with global issues—however, has broader utility for the field. The real promise for many areas of sociology, I contend, rests at the nexus of new computational methods and more traditional qualitative techniques.

A BRIEF HISTORY OF TEXT ANALYSIS WITHIN SOCIOLOGY

For sixty years, scholars have wrestled with various ways to formalize text analysis computationally (Pennebaker and Chung 2009). Whether attempting to organize and/or analyze content or to examine aspects of style, scholars have shown that computational text analysis has promising academic and commercial applications. These methods, however, by no means supplant non-computational, interpretivist approaches that preceded them and that persist. Indeed, historical, literary, and sociological methods of qualitative text analysis remain viable and important analytic techniques.[2] In fact, the most fruitful direction for the incorporation of computational methods into sociology likely demands some combination of computational and well-established qualitative techniques. Formal analytic methodologies, as Mohr (1998:346)

[2] Interpretive sociology is derived from, but not limited to Weber's historical method, which has been defined as "simply that we should try to understand the ideas and intentions of historical actors rather than search for historical laws of social evolution, as Marx and other evolutionists had done" (Roth 1976:316).

describes, can "reduce complex collections of cultural data to simpler, more easily intelligible structures of meaning." Formalization becomes a necessary step, with little or no *a priori* knowledge about what is "important." However, reducing texts to its most basic parts, such as words, often leaves out the most sociologically relevant aspects of content—content that can be brought back in after the corpus is formally "organized." While critics of formal text analysis rightly worry about the seduction of computational approaches (see Biernacki 2012 and Reed and Alexander 2009), I argue that the possibilities provided by network text analysis and other formal techniques, especially when used along with more traditional strategies, presents exciting opportunities for expanding sociological understanding.

Table 1 summarizes the predominant text-based techniques used by sociologists. From more interpretivist to more computational techniques we can see important trade-offs between breadth and detail of analyses. These strengths and weaknesses trace well-worn divides within sociology from the qualitative to the formal and quantitative, but also underscore the heterogeneity of sociological practice. Such heterogeneity warrants further discussion.

<Table 1 about here>

Socio-historical Text Analysis

Non-computational techniques consist of a series of analyst-centric methods. Centuries-old debates in literary criticism have highlighted various issues that range from ways to view text (i.e., as isolated and independent cultural objects to highly situated social constructions), the role of the author, the role of the critic or scholar, and so forth. Debates in these regards often translate into various approaches to historical analysis as well: who writes history, what are the roles of subordinate groups and official documents.

Within sociology, interpretivist approaches to text analysis are the most analogous to historical and literary methods. Scholars using these hermeneutical methods dig deep into historical events through close analyses of "social texts" (see Alexander and Smith 2003). For example, in his work on the civil rights movement and communication, Alexander (2006) describes the relationship between the messages that civil rights leaders, especially Martin Luther King Jr., delivered and their receipt by national news organizations. He illuminates the connection between the civil rights movement and the press by delving deep into the historical record, including attention to comments by the editors of large newspapers, speeches by King and others, and reported reactions by movement members in attendance. Such a detailed account is necessary to uncover the rarely explicit role that the press played in extending the activists message beyond local Southern officials and toward a broader and more persuadable audience.

Steinmetz (2008) offers a similar persuasive reading of German records in his analysis of 19th century colonialism. Developing the notion of ethnographic capital, Steinmetz describes how elite colonists constructed unique pictures of the colonized that subsequently affected how the colony was treated by the leadership in Germany. He uses official documents and letters to describe differences in ethnographic capital and its effect. There are, of course, many other good examples of how particular empirical insights and theoretical advancements could not have been developed without a close reading of historical texts. When one is concerned with issues pertaining to voice and communication, close historical readings are undoubtedly an essential aspect of a comprehensive analytic strategy.

More traditional strategies of text analysis are nevertheless not without weaknesses. Hermeneutical approaches rely heavily on decisions made by the analyst. In fact, it is the analyst's ability to decipher elusive meanings in intricate and often lengthy sets of data that

provide the payoffs of this approach. It is difficult, however, to divorce an analyst's embeddedness in a culture, both scholarly and more broadly, from the meaning drawn from text. This problematic, of course, extends to more formalized types of analysis, yet can be easier to evaluate since formalist approaches contain logics that can be replicated. Similarly, relying on breadth of knowledge as a key evaluative characteristic, interpretivist approaches can be extraordinarily time-consuming. A different set of texts can lead to very different conclusions, leading interpretivist scholars to try and gain an understanding of a complete corpus of texts rather than sampling.

Content Analysis

Content analysis bridges the divide between traditional historical techniques and more recent advances in computational text analysis. While less emergent than computational text analysis, content analysis closely interrogates text much like a survey researcher interrogates a respondent. Content analysis consists of a set of techniques that involve coding a set of variables (demographic characteristics, events, and so forth) from a set of existing documents. Hodson (1999) describes three major steps in the research process that help situate the content analysts approach. First, the analyst systematically selects a population of cases or texts. Second, the analyst constructs a coding device based on major concepts in the previous literature. Last, the analyst constructs measures of reliability and bias through mechanisms, such as inter-coder reliability (or the extent to which coders select the same score on a variable), that lend confidence to the research findings.

Content analyses have grown in popularity in the social sciences since their initial development by Harold Lasswell among others in the mid-20th century (see Mohr and Bogdanov

2013). Using data from newspapers (e.g. Walker, Martin, and McCarthy 2008) to ethnographic corpora (e.g. Hodson 2004), content analyses has clear-cut strengths. Chamberlain et al. (2008), for example, use workplace ethnographies to uncover the nuanced relationship between forms of sexual harassment and organizations. With data from over 200 workplace ethnographies, the authors uncover three distinct types of harassment that vary by workplace organizational attributes, such as grievance procedures and job instability. Obviously, such comparison across organizational forms would be difficult, if not impossible, for a workplace ethnographer to make in any systematic way. We should, however, be open to the possibility that an ambitious qualitative text analyst could read this population of ethnographies and reach similar conclusions. One key advantage of content analysis nevertheless lies in its ability to perform reliability and bias checks. For instance, in their study, Chamberlain and her colleagues are able to present high inter-coder reliability and a high degree of agreement between researchers (2008:272).

Critics of content analysis focus primarily on what is lost during the coding process. For example, coding most often consists of an *a priori* set of decisions. Content analysts must make determinations, sometimes subtle, about the presence or absence of particular variables in texts using a relatively inflexible coding strategy. More importantly, perhaps, the development of the coding scheme itself, although constructed based upon previous research, remains closely wedded to the interests and perspectives introduced by the analyst (Mohr and Bogdanov 2013). In this way, content analysis continues to walk the line between data-driven and analyst-driven research. Recent developments in narrative and lexical analysis attempt to oscillate more smoothly between emergent qualities of text data and the interpretation of these qualities.

Narrative Analytic Techniques

Narrative analytic techniques describe more emergent properties of text by using computational techniques, based on either grammar or network analysis, to allow texts to "speak" for themselves. From a grammar or network driven perspective, these techniques focus on events, and their ordering, that structure stories. By accumulating several different versions of events, narrative analysts, like others interested in narrative discourse, seek to offer "a coherent chronological story of 'what actually happened'" (Hall 1992:168). Narrative techniques rely heavily on decisions made by the analyst, but, much less so than the previously discussed techniques.[3]

In a description of his path-breaking grammar-based narrative techniques and related research that employs them, Roberto Franzosi (2004) does not abandon content coding altogether. Rather, he uses the grammatical structure of texts to narrow the lens of analysis. With this technique, the analyst does not need a lengthy coding sheet, nor does the analyst need a more holistic reading of the corpus. The analyst instead uses computer software, specifically relational database programs, to focus their attention for subsequent coding based on grammatical units, or story grammars, of subject-action-object (S-A-O). This S-A-O semantic triplet is entered and evaluated by the analyst, although Franzosi's program allows for some dictionary-building capabilities (2004:71). Unlike traditional forms of content coding, as Franzosi (2004:60) notes, "The coding categories of the grammar are functional, linguistic-based categories. They reflect properties of the text itself, rather than the researchers' theoretical interests." In this way, the

[3] Earlier work on sequence analysis (Abell 1993; Corsaro and Heise 1990; Griffin 1993) use formal methods to analyze analyst-coded narratives (see Abbott 1995 for a review). When derived from texts, these methods are roughly a hybrid of content analysis – the coding of events – and the more computational narrative analytic techniques – the modeling of these codes.

story grammar technique moves beyond the hypothesis-driven demands of traditional coding and moves to a more emergent analysis.[4]

In her research on 20th century policy makers in Philadelphia, Becher (2008) uses grammar-based techniques to describe variations in political storytelling and elite identity maintenance.[5] First, she identifies 13 interviews from a corpus of 145 that pertain to urban development. Next, she locates over 700 action statements and their constituent parts within the selected interviews following Franzosi's S-A-O method. She finds that characteristics of the texts themselves provide analytic insight into political decision-making as elites within particular groups privilege in-group members grammatically: political elites within her sample use specific names, for example, in describing in-group members and abstract names when describing outsiders. In her retrospective accounts of urban development, moreover, in-group members are much more likely to be the primary agents of positive change. While one may be able to gather these conclusions from a "thick" reading of the texts, Becher (2008) systematizes these findings, and this allows for both replication and comparison across political debates.

Network analytic techniques, an alternative, turn to the similarities in narrative structure in text corpora beyond narrow syntactical units. The network representation of narratives promotes the evaluation of the relationship between events. As Bearman, Faris, and Moody (1999:502) write, "[O]nly when we provide a beginning and an end to a sequence of interrelated events can we understand the meaning of an event within the sequence and, by extension, the meaning of an event sequence as a whole." In their network method (extended in Bearman and Stovel (2000) and Smith (2007)), events are connected to other events temporally indicated by

[4] The extent of the "story grammar" beyond the S-A-O triplet is dependent upon the researcher. As the number of variables in the grammar grows, the story grammar begins to look more and more like the coding device described in Hodson (1999) (see Franzosi 2004:339-342).

[5] In her interesting analytic strategy, Becher (2008) also uses some word count techniques as well. These techniques will be discussed in the next section.

narrative clauses, such that one event may be connected to another through an intermediary event.

Of course, narratives, or life stories, differ in how the narrator relays their biography. Such similarities and differences reveal important insights about how events can change future identities. For example, Smith (2007) describes how two European ethnic groups, Italian and Croatian Istrians living in New York City, overcame ethnic division that still existed between their European "cousins." As these groups encountered one another in a different social setting, the boundaries of their narrative changed. The narratives, in fact, grew less dependent on the memory of violence and strife, allowing the two groups to relay stories that bolstered reconciliation. This work and other research in the network analytic vein use the relationship between events in narratives to reveal similarities and differences in the aggregated story of particular historical phenomena. This method, therefore, allows the analyst space from the narratives – a vantage point that decreases the likelihood of analyst bias by permitting the pattern generated from the relationships between events to emerge through formal network analysis.

Narrative analytic techniques, both grammar and network-based, reinsert the importance of temporality—something recognized by interpretivist approaches though often ignored in traditional content analysis. These techniques also go a long way towards automating the coding schemes that make content analyses "resource rich," yet are not fully emergent techniques. Events are identified and coded, or, at minimum, extracted by hand. Many analyses derived from these methods consist of a relatively small sample of texts (see Franzosi 1997 for an exception). The promise of connecting these two techniques—i.e., network methods and more interpretive approaches—looms on the horizon as a means of efficiently analyzing the narrativity of hundreds, if not thousands, of texts.

AUTOMATED WORD-CENTRIC TECHNIQUES

Word-centric techniques turn from the analysis of events to the analysis of text corpora based on patterns in shared content or style. These techniques attempt to move beyond the description or accounting of events to the meaning behind them. An automated formal technique, word-centric analysis uses computational methods to identify patterns within a corpus. There are two main strains of word-centric analysis of texts: judge-based and word pattern analysis (Pennebaker, Mehl, and Niederhoffer (2003). Judge-based analysis consists of words that are assigned to broader categories by an analyst or group of analysts. For example, scholars interested in the portrayal of terrorism may construct a dictionary or vocabulary of words analogous to the concept "terrorism" (Carley 1997; Diesner and Carley 2005). On the other hand, word pattern analysis allows for the appearance of relationships to emerge without dictionary construction. While both methods are automated, word pattern analysis creates greater distance, and less bias, between analysts and texts.

Using emergent word pattern analysis, scholars in the information and science studies, such as Leydesdorff and Vaughn (2006), and He (1999), propose that studying the connection of thousands of words over dozens of texts can locate the dominant patterns in a corpus, or set. Using science-based corpora, such as scientific abstracts (Moody and Light 2006) or scientific debates (Leydesdorff and Hellsten 2006), these scholars have developed word network analysis techniques to locate dominant themes. This follows a rich tradition within linguistics and communications—a tradition that aims to locate meaning in a vast array of corpora (Osgood, May and Miron 1975; Danowski 1993). It is only recently, however, that they have been used to analyze historical texts in sociology.

These techniques take the word as the focal unit. As a "bag-of-words" technique, grammar is ignored. Each text is identified by a string of words, like DNA. These words are often limited to those that carry some "meaning," resulting in the elimination of prepositions, conjunctions, and so forth. These strings of words are compared using techniques described in detail below, allowing for the creation of thematic maps that describe the structure of the corpus. There are at least three advantages of this technique. First, the technique privileges transparency. Like other formal methods, analytical decisions are overt. Second, the technique allows for agnostic starts. One of the issues inherent to analyst-centric methods is the issue of how one "enters" the data. Which text does one begin reading or coding first, since each reading influences subsequent readings? This issue becomes especially pronounced when analyzing text corpora that are both large and under-studied. The notion of agnostic starts does not imply a position on inductive or deductive analysis. One can develop hypotheses about a corpus and use these emergent techniques to evaluate hypotheses. Finally, these techniques are relatively cheap and efficient. The analysis described below, for instance, can analyze thousands of texts in hours; something that would take hundreds of hours if read one-by-one by an analyst.

Recent applications of topic modeling in social science – an advanced lexical, "bag of words" technique –show the promise of this type of analysis. Topic models, developed by computer scientists, specifically David Blei and colleagues (Blei, Ng, and Jordan 2003), consist of Bayesian techniques to locate how words cluster into topics within large text corpora. The approach "reverse engineers" the process of writing by assuming that texts are distributions of topics, which are distributions of words (Mohr and Bogdanov 2013). McFarland et al. (2013) build on this, using thousands of scholarly abstracts as a case, and develop visualization techniques that facilitate sociological analysis of so-called big data. Proponents of these

techniques are generally less interested in moving back and forth between the texts and the aggregated solutions. Moreover, these techniques are particularly promising for datasets that have at minimum several hundred texts (Schmidt 2012).

As with the alternative methods of text analysis, lexical analysis has several weaknesses worth noting. A primary concern is the danger of de-contextualizing speech acts. Language and word-use are obviously extraordinarily complex phenomena and simplifying these phenomena to their core lexical properties removes a great amount of information. As Pennebaker, Mehl, and Niederhoffer (2003:571) note, "Virtually all text analysis programs that rely on word counts are unable to consider context, irony, sarcasm or even the multiple meanings of words." Several research strategies likely alleviate some of these concerns: First, increasing the volume of analysis both in terms of sample size and (if possible) the length of texts reduces the "noise" created by irony, context or the existence of multiple meanings. Second, the introduction of more information by identifying terms' parts-of-speech promotes a more nuanced view of text data. That being said, concerns about the "noise" within unstructured data are all but impossible to allay in their entirety. Rather, like any analytic technique, research must carefully specify the purchase gained when using a specific approach. In this case, lexical techniques create roadmaps for indexing qualitative immersion. Although the given "noise" is a concern, the key is to not stray completely or too far from the text itself.

AN ILLUSTRATION: PRESIDENTIAL INAUGURAL ADDRESSES

The process of constructing lexical networks based on word pattern analysis is multi-stage and involved. It remains accessible nonetheless especially when compared to alternative computational techniques. It also entails explicitness of decision-making much like using

techniques involving larger datasets, such as topic models (see McFarland et al. 2013). In my brief example here, I use presidential inaugural speeches to illustrate the network construction process. I focus on two networks that are generated from the word-by-text object matrix. In this matrix, individual words in the corpus are the rows and the speeches are the columns. Therefore, the cells correspond to the number of times a word appears in a given speech. Inaugural addresses provide an appropriate illustration because, like slave narratives, workplace ethnographies, and other historical corpora, the corpus of speeches is a specific genre (Coe and Newman 2011; Sigelman 1996).

Speeches by political leaders have received some attention particularly within the field of political psychology and leadership studies (e.g. Simonton 1988). This work builds on the contention that political speech, both in terms of content and in style, offers insight into shifts in mood and opinion, potential behavioral changes, and so forth. In an early example of political rhetoric analysis, for instance, McDiarmid (1937:79) finds "striking uniformity" in the use of symbols in inaugural addresses, from gestures of national identity – identified by phrases that identify the United States as the greatest democracy – to demarcations of expectation – identified by phrases expressing optimism regarding the future of the nation. More recent analyses have emphasized how presidential rhetoric, in inaugural addresses (Sigelman 1996) and more broadly (Lim 2002), has become more accessible to a wider audience.

Inaugural addresses, besides being useful as an example, are also profound symbolic moments that typically introduce a new presidential administration or a new term for an incumbent. Following often contentious elections, the president seeks to heal political wounds and to signal their hopes for the nation (Coe and Newman 2011; Sigelman 1996). The genre also has its limitations: First, inaugural addresses are only given once and, therefore, are not

necessarily representative of the entirety of a president's opinion or rhetorical style. Second, these are ritualized occasions. Inaugural addresses are constrained by the conventions embedded in official ceremony (Campbell and Jamieson 2008; Coe and Newman 2011). Nonetheless, as thousands of citizens who attend them will attest, inaugurations remain a calendar-setting moment in American civic life.

Given my emphasis in this article on analytic strategy, the following illustration does not delve as deeply as more extended historical analyses into rhetorical issues. Rather, I highlight similarities and differences across the inaugural address corpus to show how this analytic strategy operates and what is gained. For example, the networks will depict the relative stability of occasional words within the corpus, consistent with the pomp and circumstance of inaugurations. But, more substantively, the results show an across-time increase in rhetoric regarding the global place of the United States.

Data

The data itself consist of the corpus of United States' presidential inaugural speeches from 1789-2009. From George Washington's to Barack Obama's first inaugural speech, 56 speeches were delivered. These speeches range in length from the 133 word missive by the taciturn Washington on the occasion of his second inauguration to the 8,442 word speech by the Ohio Whig William Henry Harrison. These speeches also obviously diverge in quality, ranging from Abraham Lincoln's second, characterized by the *Washington Post* (2005) as "perhaps the greatest speech ever given by an American," to James Buchanan's 1857 inaugural, described by the *Post* (2005) as a "craven, simpering speech."

Constructing a Network Text Analysis

The first step in constructing a network text analysis involves several stages worthy of explanation. To be sure, several require a certain amount of statistical or computer programming knowledge. Consistent with other text analyses (see Diesner and Carley 2004; Moody and Light 2006), I (1) preprocessed each of the texts, removing extraneous information, such as symbols, numbers and html codes: only text pertaining to the speeches directly was included (i.e. bibliographic information was deleted).[6] Next, (2) each speech was tagged with a part-of-speech (POS) tagger (Toutanova and Manning 2000). POS-taggers, in this case the Stanford POS-Tagger,[7] identify the part-of-speech for every word in the text, and with high reliability. The Stanford POS-Tagger "reads" the speeches as .txt files and returns a .txt file with each word identified by its part-of-speech. "Blessing," for instance, is tagged as "blessing.nn" (the tag ".nn" signifying a noun). Additionally, I (3) removed uninformative words, such as participles and prepositions, by comparing the text to a standard stop list.[8]

I (4) also stemmed each of the words using the popular porter-stemmer algorithm (Porter 1980).[9] Thus, the nouns "blessing" and "blessings" are subsumed under the concept of "bless.nn" (the tag ".nn" signifying a noun), yet the verb "bless" remains independent. The resulting matrix is a speech by stemmed-word set. To reduce the risk of over-inflating highly common words – words used so frequently that they lose explanatory interest – I (5) weight each

[6] I used SAS 9.3 for this step, but could this type of preprocessing could reasonably be accomplished in numerous alternative programs, including R.
[7] The Stanford POS-Tagger can be found at http://nlp.stanford.edu/software/tagger.shtml.
[8] These words are uninformative from a content perspective, but, as Pennebaker, Mehl, and Niederhoffer (2003) describe, they are informative when attempting determine stylistic differences between texts and authors.
[9] I specifically use the porter-stemmer packaged in the Information Visualization CyberInfrastructure, produced by the Information Visualization Lab at Indiana University, http://iv.slis.indiana.edu. This program "reads" a column of words submitted as a .txt file and returns their stems in a column as a .txt file. Porter-stemmer programs are also available in R and SAS.

concept using the *tf* x *idf* model used by information science scholars (see Börner, Chen, and Boyack 2003; Salton and Buckley 1988):

$$W_{ik} = \frac{tf_{ik} \times \log\left(\frac{N}{n_k}\right)}{\sqrt{\sum_{j=1}^{T}\left(tf_{ij}\right)^2 \times \log\left(\frac{N}{nj}\right)^2}}$$

In this formula, tf_{ik} is the term frequency for term k in document i; N is the total number of documents; n_k is the number of documents that contain the term; and j indexes the terms in the corpus as a whole.

The network text analysis consists of two networks derived from this speech by weighted word matrix.[10] The goal for each is to create a functional visualization that helps tell sociohistorical stories. In other words, the goal is to move from words to network and back. With this in mind, the first analysis I show is a speech-to-speech network where the nodes (or points) represent speeches and the edges (or lines) represent the word correlation between the speeches. While some debate exists over the appropriate measure for creating similarity scores (Klavans and Boyack 2006), due to computational ease and given the similarities in results when using other measures, I use Pearson's Product Moment Correlation Coefficient. Thus, if the content of two speeches is highly correlated, they will be close together within the network. The final network, rendered in *Pajek* (Nooy, Mrvar, and Batagelj 2005), consists of the most robust component with the highest correlations between speeches.[11] This network indicates differences within the context of the corpus and provides initial evidence of the similarities and differences between speeches. Speeches at the center of this network share commonalities with many other texts, while those on the outskirts will be more dissimilar.

[10] Breiger (1974) introduced the logic and matrix algebra behind these transformations to social network analysis. Recent network techniques build upon the untransformed two-mode network, but, create intricate visualizations that are difficult to comprehend and are used primarily for measurement purposes.

[11] To find this component, I lowered the tie-strength until the weakest connected address was removed from the network and selected the penultimate tie-strength (r=.1).

The second network consists of the inverse of the speech-to-speech network. Here, the nodes are words and the edges are the correlation between speeches based on overlapping content. If words appear together in different speeches they will be closer together within the network. This network represents an overview of the content within these speeches. Clusters within this network indicate predominant themes within the corpus. As this network consists of thousands of words, I look specifically at prominent and correlated concepts (n=147). I reduce the initial 8,000 word network to those words used at least 30 times in the corpus. As these words appear in many, if not most, of the speeches, I set a lower-bound for tie strength r=.45. I use the popular Louvain method for identifying communities of words within the network (Blondel et al. 2008). Louvain method identifies nodes (here, words) that are more likely to be connected to one another.[12] The resulting network, rendered in *Pajek* (Nooy, Mrvar, and Batagelj 2005), represents the backbone of the lexical network.[13]

The Inaugural Address Network

Figure 1 presents the speech network for the inaugural address corpus. In this network the nodes represent each of the 56 speeches and the edges represent the correlation between the speeches based on overlapping content. This network represents the most robust component with the highest correlations between speeches. An important indication of face validity, the speeches of presidents who were elected more than once are graphed close to one another with the vast majority directly linked, such as George W. Bush's two addresses in the right portion of the

[12] The Louvain method is implementable in Pajek.
[13] Strategies for considering cut-points are certainly worthy of continued exploration. Modest adjustments to the node or edge reduction strategies do not substantively alter results. However, dramatically shifting the lower-bounds for inclusion of words or edges will result in significantly denser graphs. The fleshy mass of a more inclusive network (a network with more nodes and more edges) is impossible to interpret via visualization.

network or McKinley's two speeches in the left.[14] We would anticipate this level of connection as speeches by the same president are more likely to share common rhetoric and ideas.

<Figure 1 about here>

The most striking feature of this network is that it is structured by two main clusters. These clusters fall largely along temporal lines with the cluster on the right representing 20th century texts and a largely 19th century cluster on the left. Interestingly, Calvin Coolidge's 1925 speech is the exception.[15] Coolidge, widely recognized as a mediocre president (see Gilbert 1988), is tied to James Monroe in this network based on the content overlap of Coolidge's address with Monroe's 1821 second Inaugural, which is somewhat vexing because of the century that separates the two speeches.

A brief reading of the speeches themselves reveals several similarities. First, both presidents address the nation's emergence from recent wars. They both speak of the debt procured during these conflicts and the cessation of war-time funding. They also mention strategies of funding future defense-related expenditures. Moreover, each president addresses the prospects of future conflict globally with specific attention to the centuries' long "Old World" conflicts. On future European strife, both presidents proclaim American neutrality. In 1821 Monroe stated, "With every power we are in perfect amity, and it is our interest to remain so if it be practicable on just conditions." Coolidge echoes Monroe in his 1925 speech:

> It seems altogether probable that we can contribute most to these important objects [e.g. peace] by maintaining our position of political detachment and independence. We are not identified with any Old World interests. This position should be made more and more clear in our relations with all foreign countries. We are at peace with all of them.

[14] Jefferson, Lincoln, and Grant's speeches are the exceptions and are only two-steps distant (i.e. one other speech links the first address to the second).
[15] McKinley's presidency bridges the 19th and 20th centuries.

While Coolidge's placement among the 19th Century addresses is somewhat puzzling on the surface, examination of his tie to Monroe's second Inaugural offers insight into the connection between these two presidential speeches and highlights the similarities in both circumstances (i.e. the country recently engaged in war with European powers) and international policy (i.e. American neutrality to future "Old World" conflicts). Coolidge's reticence towards American engagement in global conflict resonates more with 19th century perspectives. This trend becomes even more pronounced after the start of World War II, as we will see in the discussion of the inaugural address word network.

Beyond Coolidge's unique placement in the 19th century cluster of addresses, the centrality of speeches within this network offers important lessons about this unique genre of texts. Every new inaugural cycle begins a discussion about previous inaugural addresses. Second inaugurals have the benefit of being compared to the first and often the fringe benefit of lowered expectations. But even so, with each new inaugural address, pundits and journalists reflect upon those addresses that have most captured historical and journalistic attention: These are the speeches most discussed in high school history classes and likely memorized by presidential speech writers. Prior to George W. Bush's second inauguration, for example, the *Washington Post* ranked the top 10 inaugural addresses. The top-5 were Lincoln's second, FDR's first, Teddy Roosevelt's, Ronald Reagan's first, and Truman's. With the exception, perhaps, of Kennedy's, Jefferson's first, and Lincoln's first,[16] these inaugural addresses are widely considered among the best. However, the "best" speeches are not necessarily the most central within this network. In fact, we can see in Figure 1 that none of these speeches are among the most central.[17] Rather, the

[16] Each of these speeches was in the *Washington Post*'s top-ten (2005).

[17] The sizes of the nodes within this graph vary by their betweenness centrality. Betweenness centrality captures the extent to which a node connects other nodes to the whole and is, therefore, considered a measure of importance. Within communication networks, for example, betweenness centrality scores are indicative of the amount of control

most central speeches are those that share the most in common with all other speeches. Therefore, the most central speeches are those touching most frequently on the content that is connecting the graph as a whole, such as Garfield's address and Reagan's second inaugural. While memorable speeches could have risen to the top in this regard by using the most common terms within the corpus, it may be no surprise that exceptional speeches, often written in the context of exceptional times, do not necessarily best reflect the genre as a whole.

The Inaugural Address Co-Word Network

Figure 2 presents the co-word network for the inaugural addresses. Again, in this network the nodes represent the 147 most prominent and correlated words. Edge strength is determined by the correlation of the words based on overlapping speeches. This network represents the backbone of the larger word-to-word network. This graph has two major clusters as well: While the cause of this division is less pronounced than the speech-to-speech network, the cluster on the left of the network appears to capture somewhat more philosophical concepts than the cluster on the right. This suggests two significant and split thematic modes within the inaugural address corpus worthy of closer examination.

<Figure 2 about here>

Figure 3 is a zoomed image of the left-hand sector. We can see that many of the words in this close-up are affiliated with comparatively philosophical concepts. A small cluster at the top of the zoomed sector – the economic security cluster – connects to the graph through the adjective "economic." The yellow-coded cluster contains words associated with democratic

---

nodes have relative to one another (Wasserman and Faust 1994). Within the speech network, it is indicative of "likeness" as the edges in the graph are based on content correlation. Speeches with high betweenness centrality are, therefore, more likely to either connect unique clusters of the network or connect dense collections of nodes (e.g. speeches).

philosophy: "nation," "free" and "spirit." The large central cluster – coded white – is dominated by terms capturing global issues, such as "world," "earth," "mankind," and "peace."

<Figure 3 about here>

On the other hand, as can be seen in the zoom of the right-hand sector of the word network (figure 4), terms within the inaugural address corpus also cluster around somewhat more pragmatic concerns. For example, this portion of the graph contains a tight cluster of terms in the lower quadrant specific to the occasion, such as "confidence," "office," "and oath." This sector also contains a clear policy cluster – coded blue – with terms such as "law," "policy" and "effort." The center of this sector contains word clusters capturing aspects of citizenship and civics. For example, the red-coded civics cluster is top-loaded with the words "nation," "right," and "congress," while the orange-coded citizenship is top-loaded with the words "people," "country," and "government." A high degree of overlap exists between these two clusters.

<Figure 4 about here>

The clusters identified by the Louvain method provide an emergent code for analyzing the addresses: The maps function as guides for qualitative immersion. We can begin this process by analyzing changes to the inaugural rhetoric over time. Figure 5 presents several trends from the data based on the word clusters. Most of the clusters have only experienced modest change. For example, a modest increase over time in the use of words associated with the democratic philosophy cluster, while a modest decline in words associated with the civics cluster. Following expectation, however, the cluster identifying correlated words associated with global issues has grown in use over time. This increase becomes particularly pronounced during Franklin D. Roosevelt's third inaugural. The latter half of the 20th century, of course, was dominated by

global conflict and the presidents varied less over whether and more about the extent to which the United States should intervene.

<Figure 5 about here>

Again, these clusters of words serve as emergent codes, like content coding, for more immersive analysis. In fact, one of the primary advantages of network text analysis is the ability to move back and forth between the emergent network maps and the texts. In the case of the inaugural address corpus, we can use the clusters and their locations to discuss, for example, how specific presidents engage predominant themes to describe their vision and hopes for their presidency. Here, I illustrate the value of these techniques by reading across inaugural addresses guided by the word network map and the concomitant thematic clusters.

George Washington's first two addresses are heavily associated with the citizenship and occasional thematic clusters. While the first inaugural address contains several references to the divine, they are noticeably oblique. Washington is less oblique about describing what he is tasked to accomplish. He outlines characteristics of the new democracy with specific attention to the role of the presidency. He also begins to develop the language of the ceremony speaking variously about his "call" and "duty" to serve the country both during the Revolutionary War and as president. His brief second address solely embraces these two themes with special attention to the language of the occasion forever imprinting the official language of the moment: "Previous to the execution of any official act of the President the Constitution requires an oath of office."

The early inaugural addresses rely heavily on the precedent that Washington set by continuing to define the role of the presidency and even the role of the occasion itself. As early as the 1801 inaugural, Jefferson outlines how the occasion is meant to heal the wounds following

a “contest of opinion.” The healing of wounds is a broader theme within the inaugural speech corpus and the ceremony itself is a symbol of civic unity (Sigelman 1996).

While the occasional rhetoric has remained consistent – presidents have continued to describe their “call” and “duty” and to take the “oath” – as seen in figure 5, global issues experienced a dramatic rise in the rhetoric of inaugural addresses by 1941. Franklin Delano Roosevelt’s third inaugural dramatically introduces this new global age. Speaking less than a year prior to the attacks on Pearl Harbor, Roosevelt compares the efforts towards domestic conciliation by Americans during Washington’s and Lincoln’s administrations to the importance of international vigilance during his administration: “In this day the task of the people is to save that Nation and its institutions from disruption from without.” By 1945, global engagement becomes more solidified by war and is reflected in Roosevelt’s condensed fourth inaugural. He states that the people of the United States cannot live as “ostriches” and that the collective “we” has “learned to be citizens of the world, members of the human community.”

By President Barack Obama’s first inaugural in 2009, global issues had become a core component of many addresses. The majority of the prominent words in Obama’s speech, for example “world,” “hope,” and “peace,” are identified within the global issues cluster. Within other contemporary addresses, however, these words identified by cluster 6 are even more common – in his second inaugural George W. Bush used the term “hope” twice as often as Obama and used the word “freedom” an astonishing 27 times. Unlike many of the other speeches, Obama does not return to the same term, like Bush’s use of “freedom” or William Henry Harrison’s use of the term “power” over 40 times, providing cursory indication of the relative range of Obama’s message in his first inaugural. If we briefly move through his speech

with the thematic structure of the address corpus as a guide we can see more concrete evidence of this range.

Typical of this particular genre of oration as indicated by the small cluster at the top of figure 4, Obama begins his speech with the perfunctory remarks specific to the occasion: thanking his predecessor and summarizing the importance of the presidential oath of office. Like others before him, Obama continues by making several overarching statements about the current state of American affairs discussing foreign affairs, education, and business in terms relatively consistent with the more concrete thematic strains. As he will do throughout the speech he next pivots into the more philosophical thematic arena discussing a perceived lack of confidence in the America's future. He ends this pivot with a religious image asking that "Americans choose our better future," which consists of returning to "the God-given promise that all are equal, all are free, and all deserve a chance to pursue their full measure of happiness." Obama continues this pivoting throughout the speech. For example, in the middle of the speech, he connects the notion of economic prosperity with the abstract "common good." As he begins to conclude, he connects the work of the government with the following decidedly more abstract notion: "[It] is ultimately the faith and determination of the American people upon which this nation relies."

Despite these pivots and like other president's before him, Obama leans heavily on the abstract themes of promise and hope embedded in American morality and democracy relying on the symbolic, such as his citing George Washington's speech at Valley Forge. Consistent with the modern presidency, this vision of hope is not limited to domestic issues. In fact, Obama claims that a "new era of responsibility" is required because Americans have duties "to ourselves, our nation, and the world." In the end, Obama embraces the global rhetoric of the modern presidents before him.

CONCLUSION

Text analysis has grown increasingly more popular within the social sciences. Sociology is no exception. From clear methodological statements in the hermeneutic vein (e.g. Alexander and Smith 2003) to content analyses and narrative analytic techniques (e.g. Franzosi 2004) to lexical-based analyses (e.g. Simonton 1988), previous research illustrates a variety of strategies. Recent efforts that involve taking advantage of the digitization of text in conjunction with more traditional approaches, such as hermeneutics or content analysis, offer exciting new possibilities for the systematic consideration of hundreds of texts with thousands of pages. The case of presidential inaugural addresses illustrates one strategy for this incorporation, highlighting how the word network of a corpus can identify common themes that can be used as emergent content codes for a more qualitative discussion of the addresses.

Several limitations are worth noting. First, in the "bag of words" technique described here, subtle uses of language are difficult to capture. This necessitates complementary methods for a "thick" analysis. Secondly, words are drawn to the *most* similar alter words and are not allowed to flow between possible sets of like alters. While words pulled equally between clusters of words will come to rest in between clusters, the vast majority of words will be pulled in one direction or another potentially exaggerating the sectarian character of words. Developments in the computer sciences, such as latent dirichlet allocation (Blei, Ng, and Jordan 2003), account for the possibility that some words play a role in multiple thematic clusters.

Recent and quite prominent discussions have shone a light on text analysis and the systematic analysis of large streams of data. Critiquing the slow progress social scientists have made in constructing a "computational social science," Lazer et al. (2009), for example, write, "a computational social science is emerging that leverages the capacity to collect and analyze data

with an unprecedented breadth and depth and scale." They warn that private industry and government agencies have been far more aggressive in this direction and computational science could become "the exclusive domain" of these companies and agencies. These worries are very real as the best social science from Facebook, for example, is beginning to emanate from Facebook itself or those scholars with privileged access.

Researchers from Harvard and MIT, in conjunction with the Google Books Team, recently garnered attention for an article published in January in *Science* on the promise of computational approaches to culture (Michel et al. 2011). Drawing a sample of 4 million books from Google's digitization project, the authors provide numerous pictures of 1-gram (or one-word) usage. They do not dig deeper, so that they can release the data "in light of copyright constraints" emphasizing the fact, of course, that they have access to the data illustrating the kind privileged access Lazer et al. (2009) lament.

Sociology is well-poised to lead the way and dig deeper than simple word counts, and to tell in depth stories of cultural and social change through the quantitative analysis of text. With a long disciplinary history of (occasionally uncomfortable) methodological eclecticism, sociologists are poised to tackle the two tasks that, in combination, are likely to comprise the best use of the results generated: the study of broad change and contextualization of this change. The very best use of computational sociology should not, and likely will not, stand in opposition to more traditional interpretivist techniques. Rather, they should be used in mutually complimentary ways.

REFERENCES

Abell, Peter. 1993. “Some Aspects of Narrative Method.” *Journal of Mathematical Sociology* 18:93-134.

Abbott, Andrew. 1995. “Sequence Anlaysis: New Methods for Old Ideas.” *Annual Review of Sociology*. 21:93-113.

Alexander, Jeffrey C. 2006. *The Civil Sphere*. New York, NY: Oxford University Press.

Alexander, Jeffrey C. and Philip Smith. 2003. “The Strong Program in Cultural Sociology.” Pp. 11-26 in *The Meanings of Social Life* by Jeffrey C. Alexander. New York, NY: Oxford University Press.

Bearman, Peter, Faris, Robert, and James Moody. 1999. “Blocking the Future: New Solutions for Old Problems in Historical Social Science.” *Social Science History* 23:501-33.

Bearman, Peter S. and Katherine Stovel. 2000. “Becoming a Nazi: A Model for Narrative Networks.” *Poetics* 27:69-90.

Becher, Debbie. 2008. “Narrating and Naming Positive Agents: Storytelling by Philadelphia Postwar Political Elite.” *Poetics* 36:72-93.

Biernacki, Richard. 2012. *Reinventing Evidence in Social Inquiry: Decoding Facts and Variables*. New York, NY: Palgrave Macmillan.

Blei, David M., Ng, Andrew Y. and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3:993-1022.

Blondel, Vincent, Guilluame, Jean-Loup, Lambiotte, Renaud,and Etienne Lefebvre. “Fast Unfolding of Communities in Large Networks.” *Journal of Statistical Mechanics: Theory and Experiment* 10:P10008.

Börner, Katy, Chen, Chaomei, and Kevin W. Boyack. 2003. “Visualizing Knowledge Domains.” *Annual Review of Information Science and Technology* 37:179-255.

Breiger, Ronald L. 1974. “The Duality of Persons through Groups.” *Social Forces* 53: 181-190.

Campbell, Karlyn Kohrs and Kathleen Hall Jamieson. 2008. *Presidents creating the presidency: Deeds done in words*. Chicago, IL: University of Chicago Press.

Carley, Kathleen M. 1997. “Network text analysis: The network position of concepts.” *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* 79–100.

Chamberlain, Lindsey Joyce, Crowley, Martha, Tope, Daniel and Randy Hodson. 2008. "Sexual Harassment in Organizational Context." *Work and Occupations* 35:262-95.

Coe, Kevin and Rico Neumann. 2011. "The Major Addresses of Modern Presidents: Parameters of a Data Set." *Presidential Studies Quarterly* 41:727-51.

Collins, Jeff, Kaufer, David, Vlachos, Pantelis, Butler, Brian and Suguru Ishizaki. 2004. "Detecting Collaborations in Text: Comparing the Authors' Rhetorical Language Choices in the Federalist Papers." *Computers and the Humanities* 38:15-36.

Corsaro, William A. and David R. Heise. 1990. "Event Structure Models from Ethnographic Data." *Sociological Methodology* 20:1-57.

Cousins, Ken and Wayne McIntosh. 2005. "More than Typewriters, More than Adding Machines: Integrating Information Technology into Political Research." *Quality and Quantity* 39:581-614.

Danowski, James A. 1993. Network analysis of message content. G.Barnett & W. Richards (eds.), Progress in communication sciences (Vol. 12, pp.197-222).Norwood, NJ: Ablex.

Diesner, Jana and Kathleen M. Carley. 2004. "Exploration of Communication Network from the Enron Email Networks."

Diesner, Jana and Kathleen M, Carley. 2005. "Revealing social structure from texts: meta matrix text analysis as a novel method for network text analysis." *Causal mapping for information systems and technology research: Approaches, advances, and illustrations* 81–108.

Earl, Jennifer, Andrew Martin, John D. McCarthy and Sarah A. Soule. 2004. "The Use of Newspapers in Studying Collective Action." *Annual Review of Sociology* 30:65–80.

The Economist. 2006. "The Future of Newspapers: "Who Killed the Newspaper." *The Economist*: 8/24/2006.

Escott , Paul D. 1979. *Slavery Remembered: A Record of Twentieth Century Slave Narratives*. Chapel Hill, NC: University of North Carolina Press.

Fitzpatrick, Kathleen. 2006. *The Anxiety of Obsolescence: The American Novel in the Age of Television*. Nashville, TN: Vanderbilt University Press.

Franzosi, Roberto. 1997. "Mobilization and Counter-Mobilization Processes: From the 'Red Years' (1919-20) to the Black Years' (1921-22) in Italy: A New Methodological Approach to the Study of Narrative Data" *Theory and Society* 26:275-304.

–. 2004. *From Words to Numbers: Narrative, Data, and Social Science*. Cambridge, UK: Cambridge University Press.

Gilbert, Robert E. 1988. “Psychological Pain and the Presidency: The Case of Calvin Coolidge.” *Political Psychology* 9:75-100.

Grafton, Anthony. 2007. “Adventures in Wonderland.” *The New Yorker* 11/7/2007. http://www.newyorker.com/online/2007/11/05/071105on_onlineonly_grafton (Downloaded on 5/1/2009).

Griffin, Larry J. 1993. “Narrative, Event Structure Analysis, and Causal Interpretation in Historical Sociology.” *American Journal of Sociology* 98:1094-1133.

Hall, John R. 1992. “Where History and Sociology Meet: Forms of Discourse and Sociohistorical Inquiry.” *Sociological Theory* 10:164-93.

He, Qin. 1999. “Knowledge Discovery through Co-Word Anlaysis.” *Library Trends* 48:133-159.

Hodson, Randy. 1999. *Analyzing Documentary Accounts*. Thousand Oaks, CA: Sage Publications.

–. 2004. “A Meta-Analysis of Workplace Ethnographies: Race, Gender, and Employee Attitudes and Behaviors.” Journal Of Contemporary Ethnography 33:4-38

Klavans, Richard and Kevin Boyack, 2006. “Identifying a Better Measure of Relatedness for Mapping Science.” *Journal for the American Society of Information Science and Technology* 57:251-63.

Lazer, D. et al. 2009. “Life in the network: the coming age of computational social science.” *Science (New York, NY)* 323(5915):721.

Leydesdorff, Loet and Iina Hellsten. 2006. “Measuring the Meaning of Words in Contexts: An Automated Analysis of Controversies about “Monarch Butterflies,” “Frankenfoods,” and “Stem Cells.” *Scientometrics* 61:231-258.

Leydesdorff, Loet and Liwen Vaughan. 2006. “Co-occurrence Matrices and Their Applications in Information Science: Extending ACA to the Web Environment.” *Journal for the American Society for Information Science & Technology* 57:1616-1628.

Lim, Elvin T. 2002. “Five Trends in Presidential Rhetoric: An Analysis of Rhetoric from George Washington to Bill Clinton.” *Presidential Studies Quarterly* 32:328-66.

Marcus, Michael P., Marcinkiewicz, Mary Ann and Beatrice Santorini. 1993. “Building a large annotated corpus of English: The Penn Treebank.” *Computational linguistics* 19(2):313–330.

McDiarmid, John. 1937. “Presidential Inaugural Addresses: A Study in Verbal Symbols.” *The Public Opinion Quarterly* 3:79-82.

McFarland, Daniel A., Ramage, Daniel, Chuang, Jason, Heer, Jeffrey, Manning, Christopher D. and Daniel Jurafsky. 2013. "Differentiating Language Usage Through Topic Models." *Poetics* 41:607-25.

Michel, J. B et al. 2011. "Quantitative analysis of culture using millions of digitized books." *science* 331(6014):176.

Mohr, John W. 1998. "Measuring Meaning Structures." *Annual Review of Sociology* 24:345-370.

Mohr, John W. and Petko Bogdanov. 2013. "Topic Models: What they are and why they matter." *Poetics* 41:545-69.

Moody, James and Ryan Light. 2006. "A View from Above: The Evolving Sociological Landscape." *The American Sociologist* 37-67-86.

Nooy, Walter de, Mrvar, Andrej and Vladimir Batagelj. 2005. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press.

Osgood, Charles E., May, William H. and Murray S. Miron. 1975. *Cross-cultural Universals of Affective Meaning*. Urbana, IL: University of Illinois Press.

Pennebaker, James W., Mehl, Matthias R. and Kate G. Niederhoffer. 2003. "Psychological Aspects of Natural Language Use: Our Words, Our Selves." *Annual Review of Psychology* 54:547-77.

Pennebaker, James W. and Cindy K. Chung. 2008. "Computerized Text Analysis of Al-Qaeda Transcripts." Pp. 453-66 in *The Content Analysis Reader*. Edited by Klaus Krippendorff and Mary Angela Bock. Thousand Oaks, CA: Sage.

Porter, M.F. 1980. "An Algorithm for Suffix Stripping." *Program: Electronic Library and Information Systems* 14:130-37.

Reed, Isaac and Jeffrey C. Alexander. 2009. *Meaning and Method: The Cultural Approach to Sociology*. Boulder, CO: Paradigm.

Roth, Guenther. 1976. "History and Sociology in the Work of Max Weber." *The British Journal of Sociology* 27:306-18.

Salton, Gerard, and Christopher Buckley. 1988. "Term-weighting approaches in automatic text retrieval." *Information processing & management* 24(5):513–523.

Schmidt, Benjamin M. 2012. "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities* 2.

Sigelman, Lee. 1996. “Presidential Inaugurals: The Modernization of a Genre.” *Political Communication* 13:81-92.

Simonton, Dean Keith. 1988. “Presidential Style: Personality, Biography, and Performance.” *Journal of Personalityand Social Psychology* 6:928-36.

Smith, Tammy. 2007. “Narrative Boundaries and the Dynamics of Ethnic Conflict and Conciliation.” *Poetics* 35:22-46.

Steinmetz, George. 2008. “The Colonial State as a Social Field: Ethnographic Capital and Native Policy in the German Overseas Empire before 1914.” *American Sociological Review* 73:589-61.

Toutanova, Kristina and Christopher D. Manning. 2000. “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger.” In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70.

Walker, Edward, Andrew W. Martin, and John D. McCarthy. 2008. "Confronting the State, the Corporation, and the Academy: The Influence of Institutional Targets on Social Movement Repertoires." *American Journal of Sociology* 114: 35-76.

Washington Post. 2005. “Inaugural Speeches.” Originally published on 1/19/2005. http://www.washingtonpost.com/wpsrv/politics/daily/graphics/inaug_bestworst_011905 html. (accessed July 2, 2009).

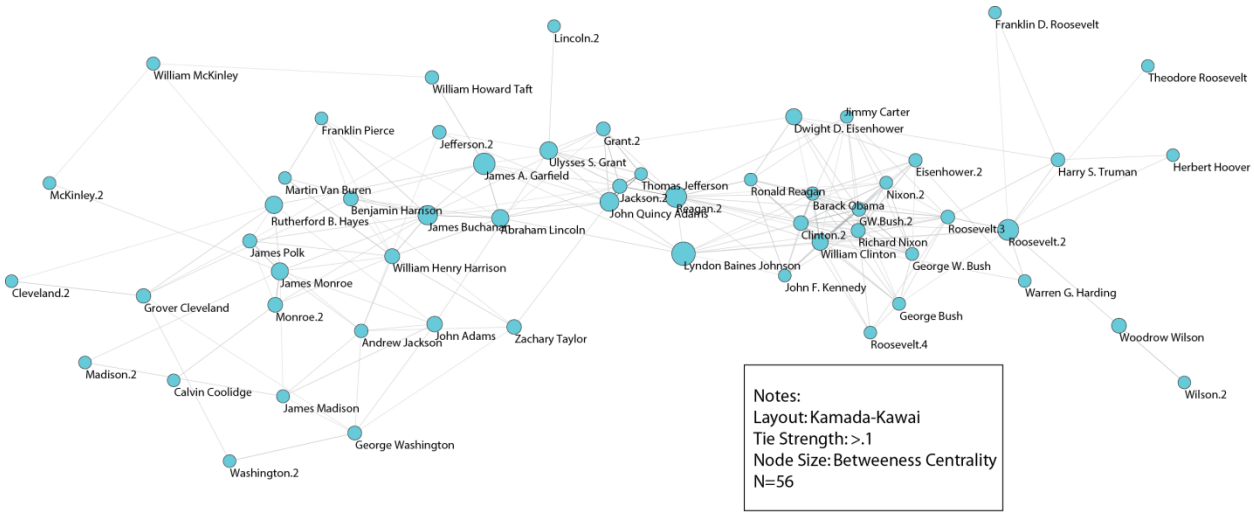
Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Young, Michael D. 1996. “Cognitive Mapping Meets Semantic Networks.” *Journal of Conflict Resolution* 40:395-414.

Table 1: Overview of Text Analysis Methods

Approach	Technique	Social Science Examples	Strengths	Weakness
Socio-Historical/Interpretivist	"Thick description," interpretivist, semiotic, etc.	Alexander (2006); Steinmetz (2008) and many others	"Depth," process, narrativity	Breadth, starting points, sampling, resources, replicability
Content Analysis	Surveying the text	Hodson (2004); Earl et al. (2004) and many others	Systematic, statistical modeling, reliability checks	Resources, depth
Narrative or Grammar-based Analysis	Grammar-based, event based modeling	Bearman, Faris, and Moody (1999); Fransozi (1997) and a few others	Time and process, hundreds of texts	Resources, black box, replicability
Word-based Networks	Lexical networks, Judge-based, Network text analysis, Topic models	Leydesdorff and Hellsten (2006); McFarland et al. (2013) and a few others	Agnostic starts, cheap, quick, intuitive, thousands of texts, replicable	Depth, interpretability issues (polysemy)

Figure 1: Presidential Inaugural Addresses Network



:

Figure 2: Presidential Inaugural Addresses Word Network

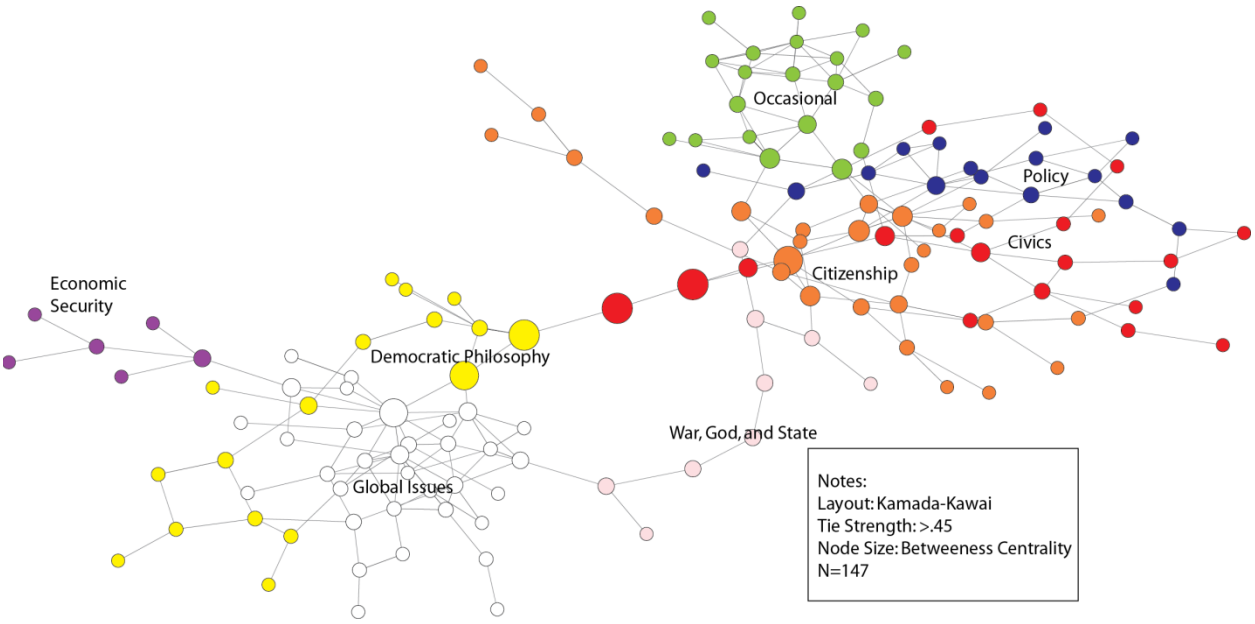
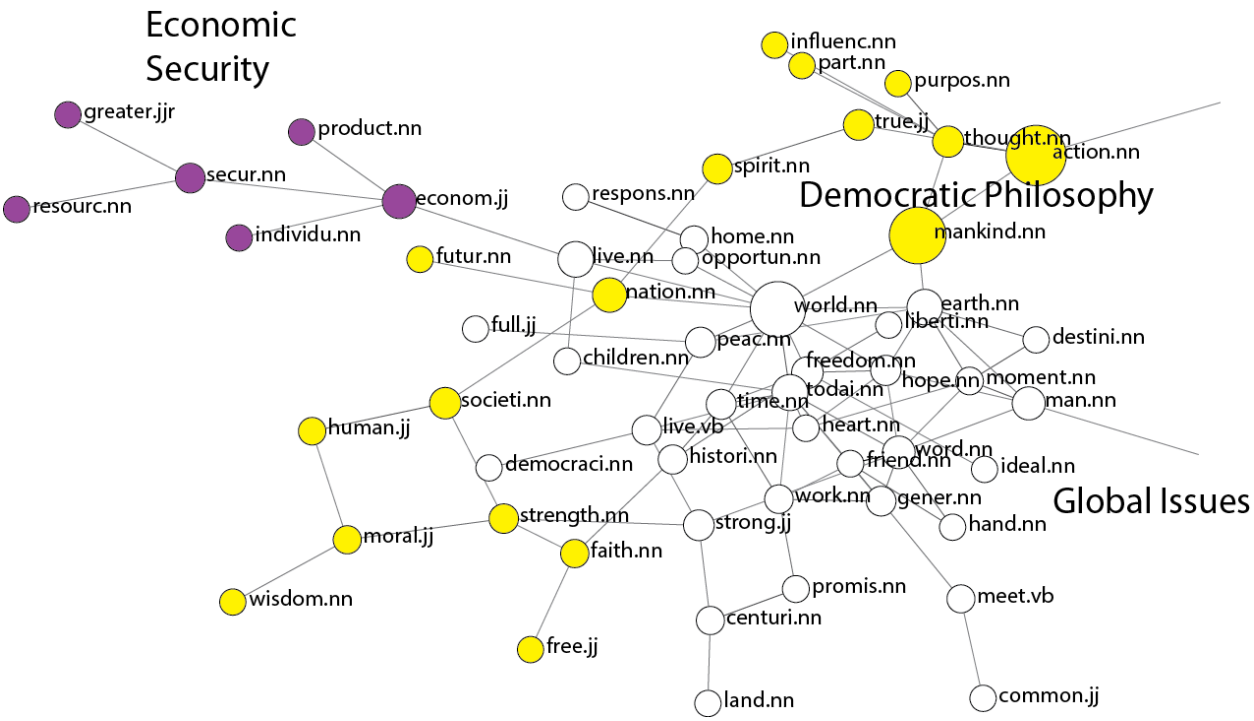
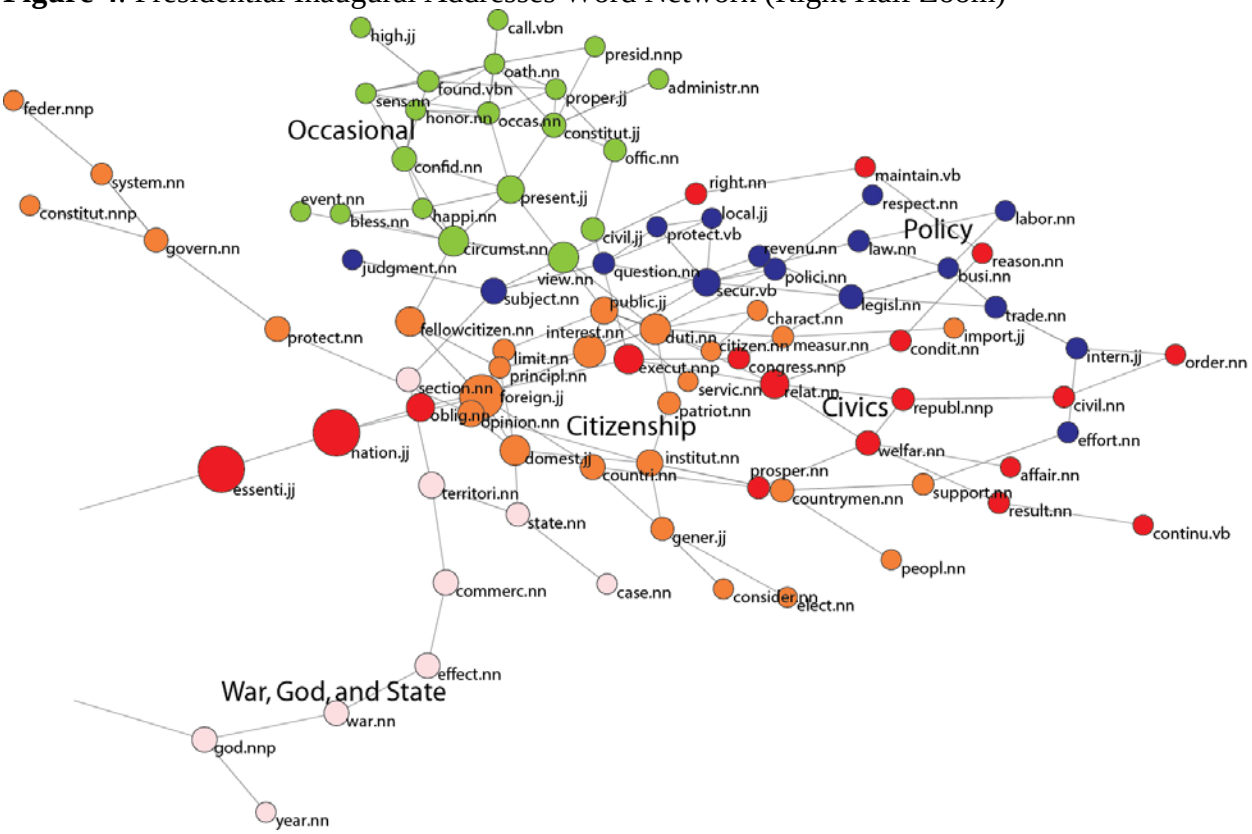


Figure 3: Presidential Inaugural Addresses Word Network (Left Half Zoom)



Tag key: .jj-adjective, .nn-noun, singular or mass, .nnp-proper noun, singular, .vb-verb, base form, .vbn-verb, past participle (see Marcus et al. 1993).

Figure 4: Presidential Inaugural Addresses Word Network (Right Half Zoom)



Tag key: .jj-adjective, .nn-noun, singular or mass, .nnp-proper noun, singular, .vb-verb, base form, .vbn-verb, past participle (see Marcus et al. 1993).

Figure 5. Trends in Word Clusters