Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014

Alix Rule[a], Jean-Philippe Cointet[b], and Peter S. Bearman[a,1]

[a]Interdisciplinary Center for Innovative Theory and Empirics (INCITE), Columbia University, New York, NY 10025; and [b]Institut National de la Recherche Agronomique–Laboratoire Interdisciplinaire Sciences Innovations Sociétés, Université Paris-Est, Marne-la-Vallée, F-77454 Marne-la-Vallée, France

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2014.

Contributed by Peter S. Bearman, June 30, 2015 (sent for review May 21, 2015; reviewed by Ronald L. Breiger and John Mohr)

This study reveals that the entry into World War I in 1917 indexed the decisive transition to the modern period in American political consciousness, ushering in new objects of political discourse, a more rapid pace of change of those objects, and a fundamental reframing of the main tasks of governance. We develop a strategy for identifying meaningful categories in textual corpora that span long historic *durées*, where terms, concepts, and language use changes. Our approach is able to account for the fluidity of discursive categories over time, and to analyze their continuity by identifying the discursive stream as the object of interest.

State of the Union | text analysis | networks | natural language processing | American history

When did modern political discourse emerge in the United States? What is distinctive of basic understandings of the tasks of governance today, in contrast to those that organized the politics of an earlier period? Can the origins of contemporary political understandings be located in the discourse of the past? The annual State of the Union address (hereafter, SoU), in which the US president reports broadly on the progress and challenges of his administration, provides a singular standpoint from which to address the evolution of the tasks of governance. It can thus be used to investigate old questions like those above using network-based text analysis strategies.

This study reveals that the entry into World War I (WWI) in 1917 indexed the decisive transition to the modern period in American political consciousness, ushering in new objects of political discourse, a more rapid pace of change of those objects, and a fundamental reframing of the main tasks of governance. At the same time, this study demonstrates that discourse distinctive to modern politics, although it later crystalized around the liberal welfare state, in fact emerged before the transition to the modern period.

We offer a unique view of American political history, which tracks the articulation of the major tasks of governance in American political and social discourse. To do so, we develop a strategy for identifying meaningful categories in textual corpora that span long historic *durées*. We are able to account for the fluidity of discursive categories over time, and to analyze their continuity by identifying the discursive stream as the object of interest. The methodological approach developed in this article can be used to meaningfully analyze texts produced over very long historical periods, where terms, concepts, and language use changes—to our knowledge, a problem not satisfactorily solved.

Historical Background

The SoU address is delivered annually by the president to a joint session of Congress, a tradition with its basis in the US Constitution, where it is mandated that the president "shall from time to time give to the Congress information of the SoU, and recommend to their Consideration such Measures as he shall judge necessary and expedient." Since George Washington's first presidential address in 1790, the SoU has been given every year, with only one exception in 1933, when incoming president Franklin Roosevelt did not give a speech. The country's first two presidents appeared in person before Congress to deliver the SoU. Thomas Jefferson, judging that this constituted an imperial gesture, set the precedent of delivering the address to the legislature in written form, a practice that endured until Woodrow Wilson took office in 1913. The latter is sometimes credited with having transformed the address into a direct appeal to the US populace, although presidents who immediately followed him sometimes reverted to written delivery. The SoU was radio broadcast for the first time in 1923, was first televised in 1947; in 1965, Johnson became the first president to cater to a television-viewing audience by delivering the speech in the evening rather than at midday (1).

Research attests to the SoU's significance in political agenda setting and the reciprocal influence of public opinion on the content of the address. The SoU reflects opinion regarding the salience of issues, while also creating it (2–4). Thanks both to its persistence and its prominence as an institution in US national politics, the SoU has been of perennial interest to researchers seeking to understand various facets of the country's history (5–9). The main focus of this work has been to pinpoint changes in political discourse to the influence of particular presidents and thus stands in contrast to the focus of this article, which is to represent continuity and change in the structure and content of American social and political thought.

To summarize, as a corpus, the text of SoU mirrors contemporary public understanding of what issues were important. It is nearly unique in the certainty and consistency of its provenance, produced at regular intervals by an individual occupying a well-defined social role, that of the US chief executive. Despite strong a priori reasons for doing so, we do not simply assume that the speech constitutes a stable cultural form, but rather demonstrate that this is the case empirically. The SoU thus provides a unique vantage point from which to reconsider arguments about the timing and nature of critical transition points in US political consciousness.

Revealing the evolution of political discourse requires appreciating how its contents change over time. The method we present

Significance

A synoptic picture of the evolution of American politics is presented, based on analysis of the corpus of presidents' State of the Union addresses, 1790–2014. The paper presents a strategy for automated text analysis that can identify meaningful categories in textual corpora that span long *durées*, where terms, concepts and language use changes, and evolution of topical structure is a priori unknown. Discourse streams identified as river networks reveal how change in contents masks continuity in the articulation of the major tasks of governance over US history.

Author contributions: A.R. and P.S.B. designed research; A.R., J.-P.C., and P.S.B. performed research; J.-P.C. contributed new reagents/analytic tools; A.R., J.-P.C., and P.S.B. analyzed data; and A.R. and P.S.B. wrote the paper.

Reviewers: R.L.B., University of Arizona; and J.M., University of California, Santa Barbara.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: psb17@columbia.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1512221112/-/DCSupplemental.

relies on the straightforward idea that words acquire meaning through their relations with other words (10). Consequently, we focus on co-occurrence, extracting the local ties between terms in paragraphs to induce categories of discourse from the resulting network structure. By recognizing that the relations between words arise in time, and appropriately defining the period over which co-occurrence is considered, we approximate the semantic standpoint of contemporary observers. We thus consider the categorical structure of discourse over successive, delimited time periods to uncover and analyze continuity and change in social and political thought. Clarifying these methodological points and identifying the insights into American social and political discourse that they permit is the focus of this article.

Methodological Background

Our analysis strategy falls into a class of text analysis methods broadly characterized as co-occurrence approaches (11), which induce categories by relying on terms' joint appearance over a particular unit of text (12). The central aim of our approach is to parsimoniously identify relevant and interpretable higher-level units of meaning endogenously, and to track their coevolution through time.

The core problem for analysts of text produced over very long historical periods is that key terms change, but for different reasons—language use shifts, new inventions join the world, concepts are recast and reorganized—making it difficult to distinguish meaningful from meaningless change. In general, canonical approaches to text analysis have not been sensitive to the fluidity of meaning over time, either on the level of individual terms or of higher-level context, conceived as categories, topics, classes, or discussions. Fig. 1 illustrates the two main reasons that a co-occurrence approach is uniquely well suited to analysis of the SoU and other historical corpora: first, in contexts where the reasons for changing word use are unclear and hard to disentangle, attention to the relationships between words is crucial for understanding the significance of such changes. Second, the co-occurrence structure, an abstraction of the changing context of use, is itself directly interpretable. In this sense, a frontal approach like co-occurrence analysis is preferable to other methods that identify categories in text, but require additional steps to make those categories accessible to interpretation.

In Fig. 1, we observe immediately that some of the terms associated with "constitution" change: "constituents" is present only in the first period, "slavery" only in the second, and "land laws" and "ideals" distinguish the semantic neighborhood of the term in the third and fourth periods, respectively. However, the relationship of these associated terms to one another also changes, strikingly. In the first decades of the country's history, "people" and the objects associated with it appear as a distinct community (colored blue), indicating one context of the constitution's meaning. During the Civil War and Reconstruction era, captured in the second period, the largely familiar set of contexts to which "constitution" is related themselves become more closely associated—as the constitution becomes central to a number of key discussions of the era. In the third period, the context of "constitution's" meaning again becomes more straightforward and more limited—its contents focused on jurisprudence—and even more so in the fourth period, although the related terms again shift. Equally, Fig. 1 allows us to see that terms that reappear successively in connection with "constitution" may undergo semantic transformation. In the first period, "confederacy" is associated with "years" and "members"—referring, in the first decades of the country's history, to the organization of member states—whereas in the second, it is associated with "state," "self-government," and "union." We do not in fact need to know that the "confederacy" was the name adopted by the seceding southern states, changes in the network of terms alone indicates that such a transformation has occurred. At the risk of being didactic, the changing significance of words—revealed by the structure of co-occurrence with other words and terms—can play havoc with traditional dictionary and topic model approaches. We expand upon the reasons for this below.

In contrast to the approach developed here, dictionary-based methods compare words observed in a corpus against a predefined and often structured set of terms. They thereby fix both a semantic structure and the definition of particular words within it. Such methods thus assume a specific substantive context (13)—for example, a dictionary for political discourse would not capture the meaning of the same terms used in everyday speech. This makes dictionaries inappropriate for corpora that span long time periods, because adopting a "substantive context" entails arbitrarily assuming a fixed historical standpoint—in our example, the political discourse of a given moment (cf. ref. 14). Supervised text classification, by contrast, builds automated classifiers inductively, which learn the characteristics of those categories they apply from a set of preclassified texts. In analyzing language that evolves over time, however, supervised text classification methods present limitations similar to those of dictionary-based approaches in that they assume that categories possess a stable textual signature.

By contrast, topic modeling comprises a set of methods for identifying meaningful categories in textual corpora endogenously. Approaches like latent Dirichlet allocation (LDA), popularized by Blei et al. (15), infer what a unit of text is about by relying on probabilistic models based on observed word distributions. For a given corpus, the analyst sets a number of topics that are then

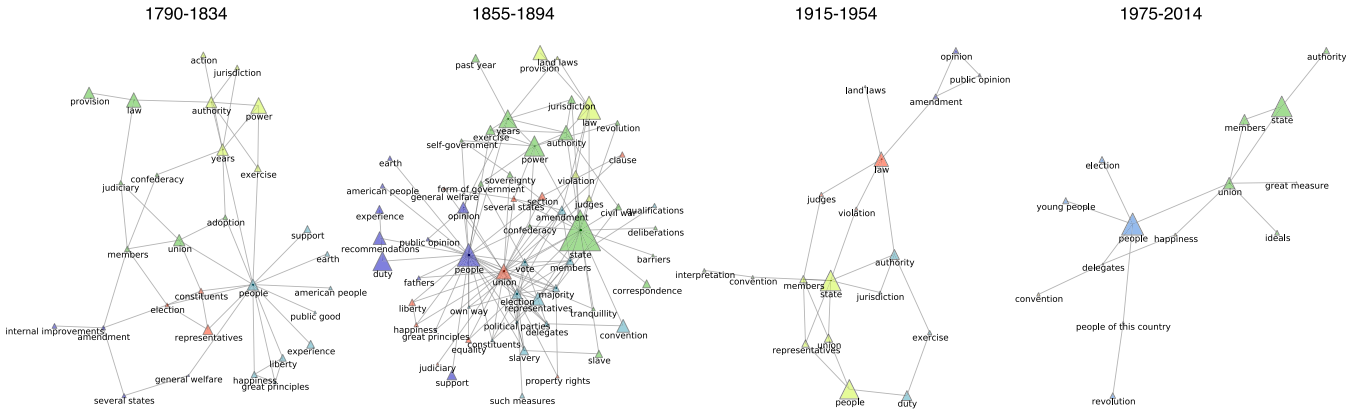

Fig. 1. The meaning of words is conditional on their co-occurrence with other words and terms. Attending to patterns of co-occurrence over time captures their evolving meaning. Key terms co-occurring with "constitution" are shown for four periods over the SoU corpus, 1790–1834, 1855–1894, 1915–1954, and 1975–2014. For each time period, we build a proximity network where each node (word or term) is linked to its closest neighbors. Nodes are colored according to the connected component or community to which they belong. The target node—"constitution"—and its links to all other pictured terms, is hidden from the visualization. Node size scales with frequency of terms' occurrence.

distinguished by probability distributions over words specific to each topic. Topics are unstructured and consist of ranked lists of terms, whose weights are associated with their likelihood of being used when a topic is drawn. A distribution of topics over units of text and the membership of terms in topics are jointly optimized.

Despite their popularity, topic models raise a number of concerns: one is that the topics they identify are not in themselves easy to interpret, and consequently not substantively "trustworthy." [The fact that topics produced through LDA often lack "actual and perceived accuracy" as meaningful categories (16, 17) has provoked topic modelers to find additional methods of checking for topics that do not fit the data (18).] Concerns about transparency further compound problems of parameterization that topic models inevitably raise for analysts (19), e.g., how many topics to specify (20).

More recent developments in the topic-modeling vein have started to account for the way that meaning arises over time. In contrast to the original probabilistic latent semantic analysis (21) and LDA approaches, some topic models now accommodate hypotheses of topic dependence—allowing for words' distributions to be correlated (22)—and for topics' dependence on a range of external variables (23), time among them (24). An example of the latter, the topics-over-time (ToT) algorithm (25) identifies topics that vary in their temporal profiles over time-stamped corpora, arguably making it well suited to detect brief semantic responses to exogenous events. However, ToT is not designed to capture how a changing set of words or terms compose a topic over time. Blei and Lafferty's (26) continuous dynamic topic modeling (cDTM) does do this. Like LDA, cDTM requires the analyst to specify a set number of topics in advance, through which different sets of words may move over time, as through a tube. Such an approach is, however, unsatisfying in cases where the genealogy of relevant categories is itself unknown, and in fact is the question of interest, as in our study.

To our knowledge, Gao et al. (27) have made the only similar attempt to address this problem, inducing clusters of documents to study the "overall evolution of topics and their critical events." However, the topical structure that Gao et al. (27) generate is not directly interpretable. A second step is needed to make topics accessible at the level of meaning; this is accomplished through an analysis of word co-occurrence "on top of" the heirarchical Dirichlet process model for retrieving the categories. In contrast to this strategy, our approach is both more parsimonious and transparent. Specifically, as illustrated in Fig. 1, our strategy exploits the co-occurrence structure—assumed absent in topic models (18)—as it unfolds, to track the continuity, discontinuity, and relations between categories across time, relying only on terms' joint appearance over a particular unit of text to endogenously induce topics.

We analyze the co-occurrence structure using familiar network analysis techniques, relying on community detection to identify and interpret categories of relevance. We then limit consideration of co-occurrence to successive time periods in SoU discourse, and describe the genealogical relationships between categories. To do this, we assess the similarity between network clusters in successive periods.

A variety of techniques for community detection in dynamic networks have recently been developed (28–31)—although they have not been widely applied to problems of semantic detection, with the exception of ref. 32. Growth in this area makes us optimistic that the approach developed here can be applied widely in the analysis of other corpora where change in terms' use over time is substantial, has multiple sources, and has uncertain bearing for higher-level meaning.

Methods

The SoU corpus includes 227 dated addresses, comprising a total of 1,763,622 words. The speeches vary in length and elegance. Our method is insensitive to such variation. We base our analysis on frequently occurring noun terms, including multiword phrases, e.g., "national security," "local government," "fellow citizens," extracted using natural language processing (NLP) techniques.

Semantic categories are induced by considering the co-occurrence patterns of terms over documents spanning particular periods of time. We define the co-occurrence matrix as the number of joint appearances of terms i and j in the same paragraph in a document published at time t. A proximity score is then computed to measure the relatedness of each pair of terms, yielding a weighted semantic network with terms as nodes connected by edges weighted by their similarity (33). A community detection algorithm (34) is used to identify cohesive subsets. These clusters are what we will refer to as discursive categories, and interpret as such. Inducing discursive categories through co-occurrence provides a rich perspective on the inner structure of each topic, illuminating the individual connections between words, the positioning of terms in the clusters, and proximity between topics.

We first apply the procedures described above to produce the entire ("global") semantic network on the full 1,000 × 1,000 terms matrix over the SoU's history. We then apply the same procedure over successive, delimited time periods of the SoU to produce historically specific semantic networks.

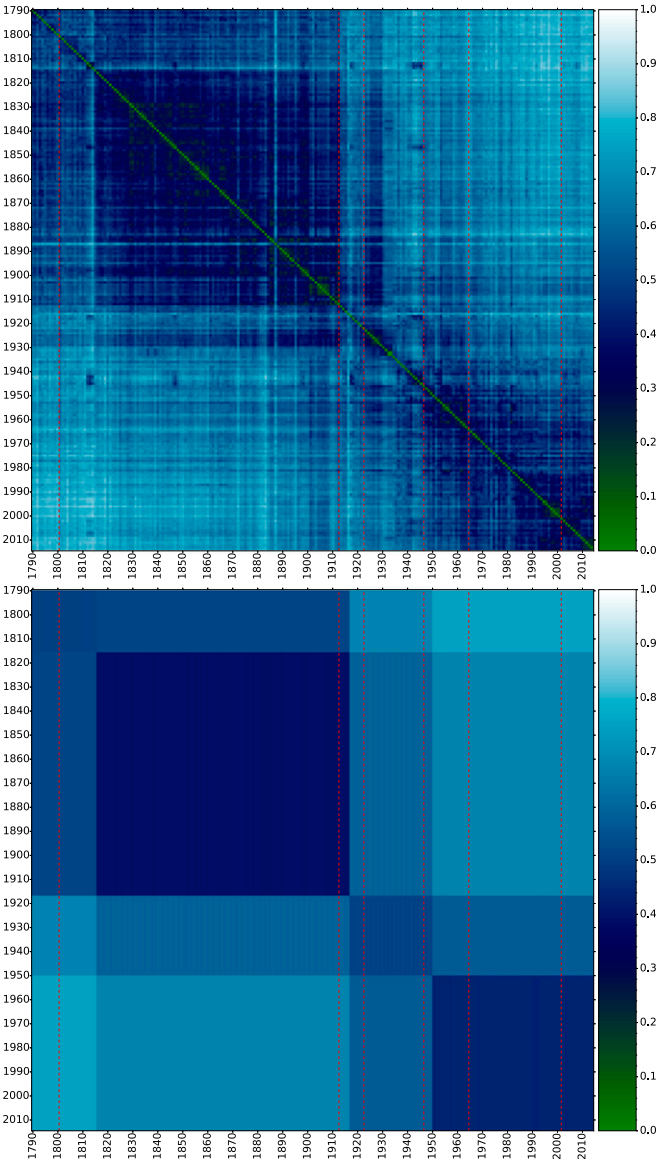

Fig. 2. The SoU is a continuous cultural form, its moments of disruption indexing real world events. The upper panel reports the transition matrix capturing change in key terms used in the SoU over time. Each cell of the matrix compares the terms vectors for two addresses, representing their dissimilarity on a scale of 0–1. Darker regions represent areas of substantial similarity; light bands represent moments in which the speech's content departed dramatically from that of (all) other years. The lower panel is a homomorphic reduction of the first panel and highlights key transition periods—1917, 1816, and 1950—none of which overlap with changes in mode of delivery (represented as red dotted vertical lines).

We refer to these as local networks, in contrast to the global network induced from the entire corpus. To analyze the continuity of discursive categories, we apply an algorithm that captures the movement of terms between discourse clusters in past and successive periods, and thereby allows us to reconstruct the most likely lineage of each discursive category. This makes it possible to represent the SoU as a series of conversation streams, in a river network.

To induce the river network, we generated local semantic networks according to the same procedure described above for 10 successive overlapping periods of 40 y each, evenly spaced to cover the entire SoU corpus. We hence obtained 10 terms maps based on the co-occurrence of the most frequent terms in each period. Clusters on these networks index historically specific categories. Each network provides an objective map to which the social and political discursive categories of contemporary actors corresponded. We then applied an algorithm (*Supporting Information*) to find the most likely lineage between the discursive categories of a given period and those that preceded it, on the basis of shared terms weighted by their within-cluster centrality. Recall that clusters index discursive categories. For each time period t, the algorithm considers a cluster detected within the given time period and knits it with clusters from the previous time period. To determine which clusters to connect, we compute the Bhattacharyya coefficient between the normalized centrality distribution of terms in a given cluster and each of the normalized centrality distributions of terms in clusters detected at $t - 1$. Pairs of clusters are intertemporally linked if similar.

Results

In terms of its lexical contents, the SoU is remarkably stable, changing only gradually over time. To assess change over the SoU's history, we first use a vector space model as an efficient tool for representing the similarity of two given addresses. We treat each speech as a vector of terms weighted by their frequency, and compare each of the dated vectors using the cosine measure, controlling for speech length. Results are displayed in the transition matrix in the upper panel of Fig. 2. Transition matrices considering 500 and 1,500 terms return results consistent with those reported here, as do an otherwise-identical set of matrices obtained using a Euclidean distance measure.

Fig. 2 reveals the gradual nature of year-to-year change in the SoU's contents, which can be seen in the dark colored cells above and below the main diagonal, indicating the similarity of adjacent years. Furthermore, one observes only a few moments of relatively brief disruption, none enduring more than a few years. Notably, the timing of these moments does not index changes in the speech's mode of delivery or broadcast, but coincides with real world disruptions: the War of 1812 and the First and Second World Wars. In short, the SoU is a stable cultural form, and dramatic departures are brief and respond to exogenous events. [The exception that proves the rule concerning the SoU's formal consistency comes in 1887, when Cleveland devoted his entire speech to a plea for tariff reform, the central issue of his campaign for reelection in the following year. Cleveland lost to his Republican challenger, which may help account for the fact that departure from the basic format of the SoU was never repeated (35).]

To analyze stability in the SoU's contents, we compared adjacent years in which the president remained in office, to adjacent years in which a new president delivered the address. Given our orientation to the changing conceptual backdrop of US politics revealed through SoU discourse, our aim was to test the assumption that the objects that feature in political speech change at a rate independent of what is said about them. Indeed, there is no statistically significant difference in the pace of change in terms either when the new president is from his successor's party or from the political opposition. In adjacent years where speeches were given by the same president, the average change ratio was 0.31.

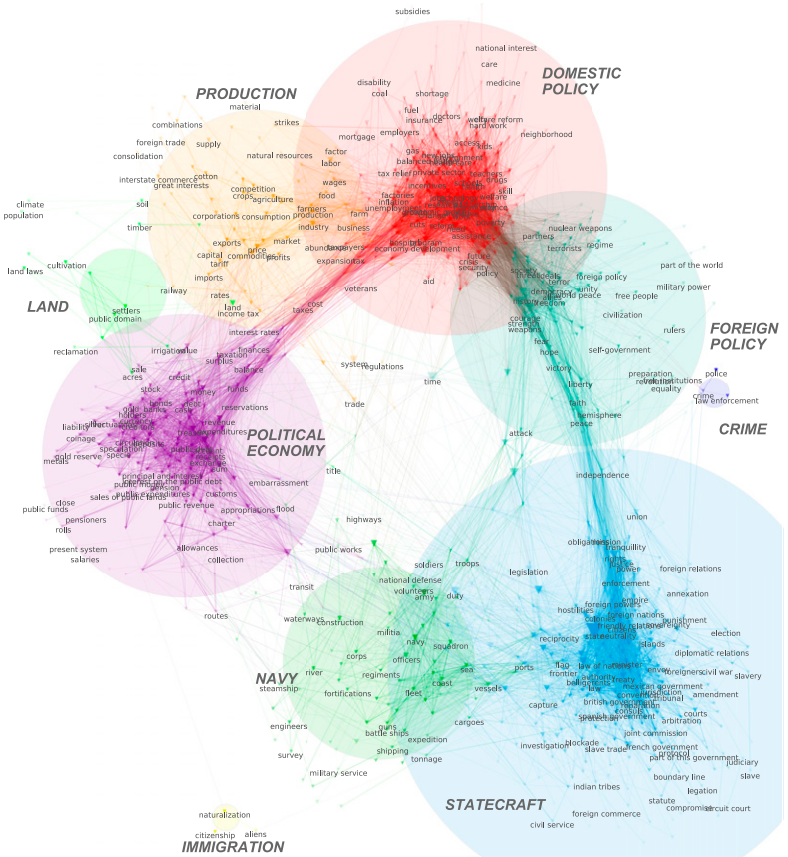

Fig. 3. The global network structure of the SoU, 1790–2014. The Louvain community detection algorithm reveals cohesive clusters or discursive categories from the semantic network built from the 1,000 × 1,000 terms matrix over the SoU's history. Some terms lie between clusters and serve as bridges connecting otherwise disjoint discourses. Two clusters contain only a few linked terms: one indexes the set of concepts associated with immigration, the other those associated with crime.

By comparison, years in which a new president from his predecessor's party gave the address averaged 0.34 ($P = 0.95$), and those in which the SoU was delivered by a new president from the opposition, 0.33 ($P = 0.85$).

The observation that the SoU is a stable cultural form, however, does not imply that we cannot describe periods of continuity and discontinuity in the objects of political concern. To detect such periods, we divide the transition matrix into subblocks that maximize the homogeneity of the average values of each pair of years falling within each block. The results of this procedure are illustrated in Fig. 2. One can immediately see that these periods of similarity are not aligned with changes in the mode of the SoU's delivery, any more than are momentary disruptions. The optimal partition is in 1917; secondary partitions in 1816 and 1950 demarcate unique discourse regimes. Differences in color saturations of blocks along the main diagonal reveal that each of these periods is also characterized by a unique rate of change. Net of further partitions, this is equally true of the history before and following 1917. The longer first period is characterized by less dissimilarity between any two given years than the second (*Supporting Information*). 1917 was the first year that objects of political discourse resembled our own more than those of the 19th century, and marks off a previous era of slower change from one a subsequent one in which change proceeded faster. It demarcates the transition to modern political consciousness.

Transhistorical Categories in US Social and Political Discourse. The bag-of-words approach above (ignoring the structure of co-occurrence) provides a baseline picture of turnover in the objects of political discourse. It thus constitutes an efficient way to detect precise moments of change. However, it is inadequate to describe what is being discussed, much less the finer contours of discursive categories or relations between them. Relying on the approach described in *Methods*, we next identify the main categories over the entire history of the SoU. Fig. 3 displays the semantic map of social and political discourse over the entire period from 1790 to 2014. Filtering the aggregated matrix of co-occurring terms across the entire corpus, results in a network with 887 nodes and 13,302 edges. The nine clusters on the network are labeled on the basis of the lexical contents as they relate to the network structure detected by the Louvain algorithm, and index the main categories of political discourse over US history. Recalling that both the internal structure of the clusters and their position in relation to one another is significant, we describe the categories captured in the terms map, clockwise from *Upper Left*: "Production" in orange is closely connected to a larger cluster broadly concerning "Domestic Policy" (red), in which economic language mingles with terminology relating to the welfare state, also connected to a discourse of "Foreign Policy" (dark green). Below this, but clearly distinct, is a discourse concerned with the functioning and external affairs of the government, "Statecraft" (blue), connected to a conversation about the "Navy" (light green). Loosely connected above is a large cluster representing "Political Economy," in which monetary policy appears as a distinct region. "Land" features above this, as a small but distinct discourse over US history.

These master categories provide a semantic summary of the entire 200-y *durée* of the SoU, but they do not reflect the native categories of speakers. No president may have ever uttered a paragraph about "statecraft" or "domestic policy." These topics are rather a meaningful abstraction of discourses historically situated individuals could have and did perceive.

Three of the seven major clusters, Land, the Navy, and Production, are largely comprised of terms that retain unique meanings over the period from 1790 to 2014. We might thus assume that, despite changing contents, these categories are semantically stable and that over the history of the SoU and that the terms in these clusters reference a single focus of substantive political concern. By contrast, four of these seven clusters, by far the largest, appear to comprise many terms that are historically multivocal. Attention to the structure of co-occurrence suggests that these categories arise from discourse conducted over specific periods of history. Namely, the Statecraft and Political Economy clusters belong to an earlier moment in history than the Foreign Policy and Domestic Policy categories. The two pairs index similar tasks of government, as constructed differently in different epochs.

Discourse Categories over Time. To obtain a provisional picture of the historical foundations of these master categories in US political discourse, we project the clusters derived from the structure of co-occurrence in the entire corpus onto paragraphs of dated SoU addresses. Specifically, each of the nine clusters is represented as a vector of member terms, weighted by centrality, and is compared with a vectorial representation of the lexical content of each paragraph. We match the content of dated paragraphs in the SoU to the clusters they most closely resemble, assigning paragraphs to one or multiple clusters. The projection procedure shares similarity with LDA topic modeling insofar as the optimal categorization of paragraphs is conditioned on topics as distributions of terms. (However, here, paragraphs' assignment to topics is accomplished in a second step and does not assume a distribution of topics over paragraphs.) The procedure allows us to track the relative representation of political master categories over time, as in Fig. 4.

The results confirm the impression of historicity of the topics represented by four large clusters on the global network. In Fig. 4, continuity and discontinuity are clearly discernible. Some clusters, Land and the Navy, largely disappear over time. Others appear to shape-shift: notably, we observe that the discourses of Foreign and Domestic Policy indeed succeed and eventually replace the discourses of Statecraft and of Political Economy. The latter two categories dominated the SoU discourse during most of the premodern period, along with discussion of Land and the Navy.

The discovery of two sets of historically distinctive categories resonates with and enhances our understanding of the 1917 transition: political discourse changed not only in its objects of concern and their pace of change, but in its construction of the basic tasks of governance. Interestingly, the results displayed in Fig. 4 suggest that the modern categories that began to dominate political discourse after WWI, first emerged before it—around the turn of the century.

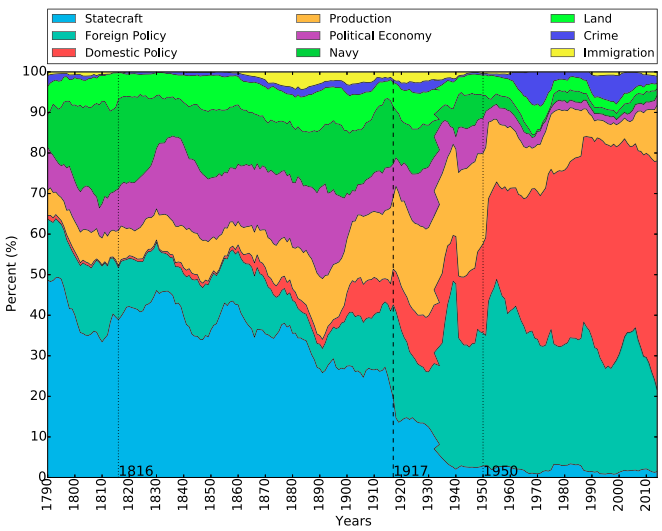

Fig. 4. Foreign Policy and Domestic Policy Supplant Statecraft and Political Economy as master categories in American political discourse over time. Projecting global network clusters onto paragraphs of dated addresses allows us to track the historical foundations of basic understandings of the tasks of government revealed in the SoU. Time (1790–2014) runs along the *x* axis, and the *y* axis reports the proportion of the SoU devoted to each discourse cluster in given year. Dotted lines demarcate periods.

Although discourses of Foreign and Domestic Policy were hardly hegemonic—neither ranks among the top three categories of SoU discourse in any year before WWI—both understandings gained ground in the first decade of the 20th century.

The implications of the projection procedure remain suggestive as to timing of semantic change, however. This is in part because the abstraction inherent in categories render them remote from those salient to contemporary observers. Furthermore, the procedure reflects the unsatisfying assumption that a single term retains the same meaning over more than 200 y. We transcend this assumption in the next section to obtain a detailed picture of the evolution of political discourse across US history.

Dynamics: River Networks. How did a different understanding of the fundamental tasks of governance emerge in American political consciousness? Did new topics of discussion appear? Were extant discussions discontinued, or reorganized? To pursue such questions, we need a way to capture the context of meaning in which modern political categories could emerge. We must reconstruct the flow of political discourse and attend to the right moment therein. To achieve this, we induce a river network.

Recall that we generated local semantic networks for successive overlapping periods, retrieving terms maps based on the co-occurrence of the most frequent terms in each. The length of the periods (40 y) reflects what could have been perceived by contemporary actors. The topics indexed by clusters on these local networks are thus are unlike the categories over the full corpus, in that they are meaningful from a particular historic standpoint, and not sensitive to semantic changes that occur in subsequent periods. We then knit these clusters together. The river network that results from this procedure captures the flow of political discourse across US history, as shown in Fig. 5. Topics (clusters) woven together across periods catenate into continuous discourse streams. Clusters so connected at t1 and t3 may comprise none of the same terms (equally, unlike in the master categories derived from clusters on the global network, then projected back onto paragraphs, particular terms may appear in different streams during different periods). A stream remains the same thing from period to period, although it need not remain one thing. The approach can recognize multiple relationships between the structures of adjacent periods. Discourse streams may fork, merge, decline, swell; new streams can always emerge and old ones disappear. Fig. 6 provides one detailed example of the forking processes that the river networks can identify; here for the transition from the cluster labeled "action and law" to the two clusters labeled "departments and recommendations" and "law and interstate commerce" over the period from 1875 to 1914, in which a moralized conversation about the administrative structure of the emergent bureaucratic state is decoupled from the regulatory structure, in this instance focused on railroads.

Two systems of interconnected streams run the full length of US history, one concerning international and the other domestic matters. For most of the country's history, discourse about fiscal policy, on one hand, and farming and industry, on the other, ran in parallel; these merged only in the mid-20th century (p8) into a unified discourse of the modern, domestic, economy. Conversely, what exists today as two distinct clusters, one concerning the United States' role as a superpower, and another about national security, both flow from common origin in a mid-20th-century (p8) discursive stream, which in turn flows from one branch of a conversation that forked for the first time in the period centered on WWI (p6) as the United States adopted an internationalist foreign policy. The other branch—the remaining discourse around bilateralism—died out within the next 20 y. Two streams cover a substantial span of the country's history. One runs from the founding period and is concerned with the country's defense infrastructure, in particular the navy; it concludes with discussion of the military of WWII (p9). The other begins in period 5

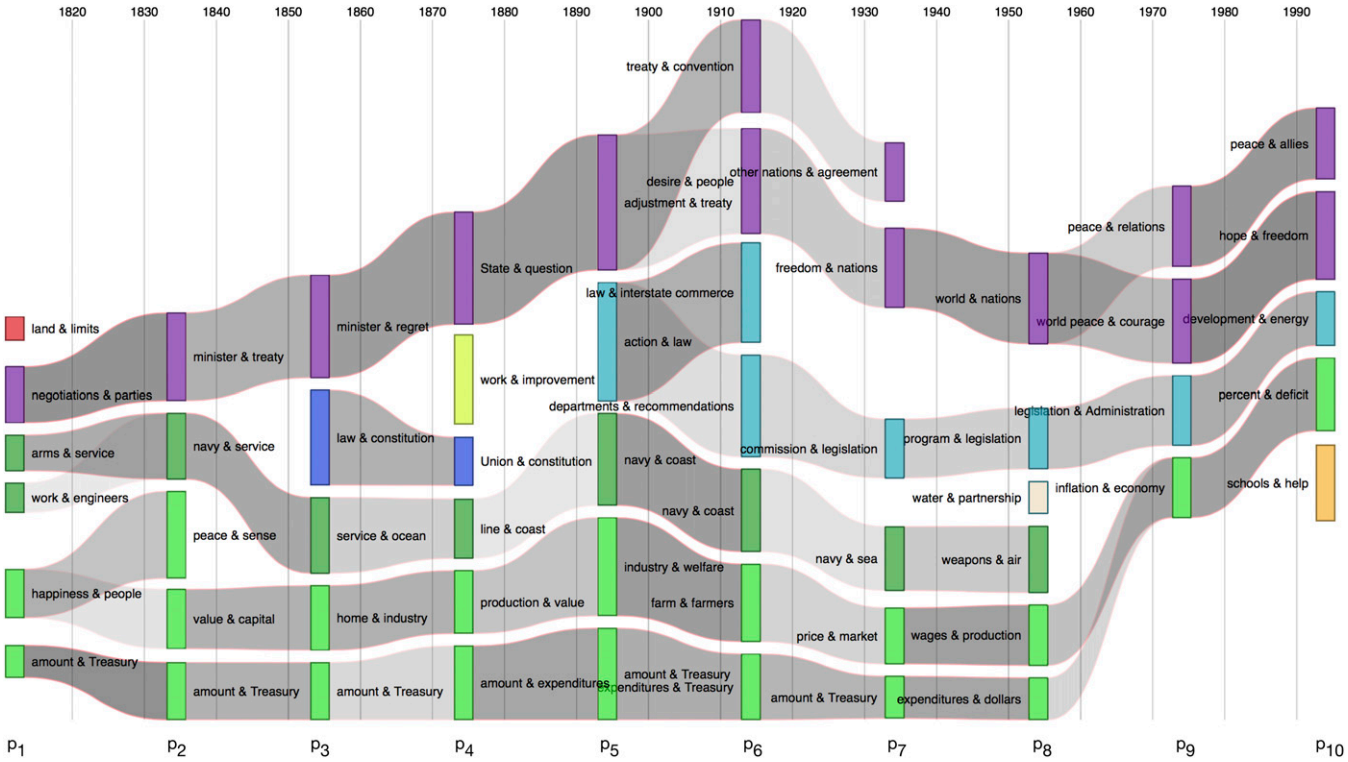

Fig. 5. A river network captures the flow across history of US political discourse, as perceived by contemporaries. Time moves along the *x* axis. Clusters on semantic networks of 300 most frequent terms for each of 10 historical periods are displayed as vertical bars. Relations between clusters of adjacent periods are indexed by gray flows, whose density reflects their degree of connection. Streams that connect at any point in history may be considered to be part of the same system, indicated with a single color.

	action & law (1875-1914)	departments and recommendations (1895-1934)	law & interstate commerce (1895-1934)
action & law (1875-1914)	liberty; justice; passage; adoption; principles; election; reform; crime; members; enactment; district; wrongs; suggestions; delay; organization	results; session; board; appointment; details; recommendations; methods; bureau; importance; attention; civil service; experience; character; Administration; system; examination; change; head; investigation; branch; establishment; necessity; information	jurisdiction; judges; railroads; decision; exercise; amendment; bill; authority; power; business; act; action; respect; legislation; corporations; courts; compensation; property; employees; regulations; persons; law; cases; railway; constitution; statute; supervision; combinations; opinion
departments and recommendations (1895-1934)	results; session; board; appointment; details; recommendations; methods; bureau; importance; attention; civil service; experience; character; Administration; system; examination; change; head; investigation; branch; establishment; necessity; information	office; committee; approval; close; last year; operation; subject; requirements; departments; commission; maintenance; veterans; purpose; consideration; extension; proposals; reports	
law & interstate commerce (1895-1934)	jurisdiction; judges; railroads; decision; exercise; amendment; bill; authority; power; business; act; action; respect; legislation; corporations; courts; compensation; property; employees; regulations; persons; law; cases; railway; constitution; statute; supervision; combinations; opinion		transportation; interstate commerce; rates; public; employers; message; account; matter

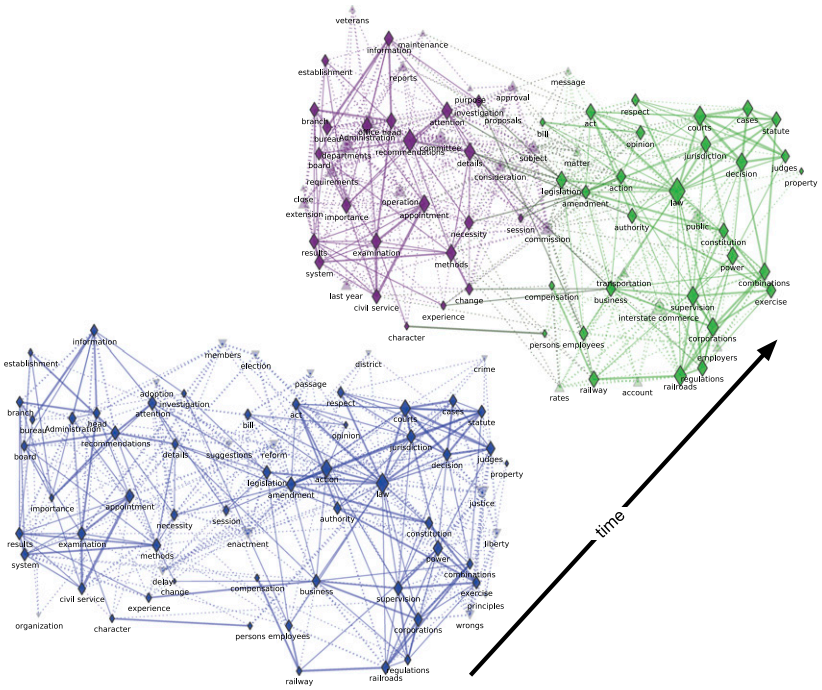

Fig. 6. The discursive categories of a given historical moment may split in subsequent periods. (*Right*) "Action and law" discourse cluster of p6, and its two successor discourse clusters in p7, "departments and recommendations" and "law and interstate commerce." Clusters' shared and unique member terms are displayed in the table at *Left*.

(1875–1914) and splits in the following period where one branch concludes, the other continuing unbroken to the present, tracking the emergence of the contemporary welfare state.

Although generated no less systematically than the static master categories, the streams recapture the unfolding of American political discourse at a high level of historic detail. Also evident in Fig. 5 are a number of local discussions, which are too specific to their moment to map easily to the master categories described above. Two are time-censored: the contemporary discussion around education and communities "schools and help" that regards market-driven social policy has not had a chance to extend into the future; likewise, the early discussion about control of territory "land and limits" was inherited from an earlier period not captured in the SoU. The stream that covers the political crisis culminating in the Civil War and the discourse of Reconstruction spans only two periods, and disappears without disrupting the overall structure. This is consonant with the results of the previous lexical analysis that demonstrated that the long 19th century (from 1816 to 1917) was a period of marked stability with respect to social and political discourse.

Streams capture qualitative continuity in topics of political discussion, but the consistency of their contents across history is uneven. We calculate the rate of new terms, terms pairs, and pairs of terms remaining in the same category of political discourse for the nine transitions between our ten periods, which gives us a comparative picture of how radically the most important objects in political discussions of the day were changing, underneath the continuities revealed in Fig. 5—and much enhancing the low resolution of change provided by the vector space model (Fig. 2). Here again, we find that the moment following WWI (the period 7–8 transition) is an outlier, exhibiting the most radical turnover in the key objects political discussion all three dimensions—but that notably this change occurs without a fundamental reorganization of salient political categories. What was being talked about changed in the political discourse of the post-WWI period, but the framing of basic issues of governance did not.

Discussion

As striking as what is captured in our analysis of the evolution of American political thought is what is missing. The Civil War, often considered in conventional histories to have transformed the country's political consciousness, while apparent in political discourse of the time, seems not to have made a lasting imprint on the unfolding of the dominant categories of social and political thought in the SoU. Although discordant with the organization of introductory textbooks, the absence of distinct periodization observed in this study around the Civil War is consistent with historical scholarship that identifies the main conversation in the antebellum period as centered on states' rights, a debate the war, and reconstruction and its collapse, failed to solve. Likewise, our study challenges histories that identify Reconstruction, the New Deal, and WWII as inaugurations of modern political consciousness in the United States. (Although we find some support for the Marshall Plan as marking a secondary transition within the modern period.)

The central finding is that the modern understanding of politics began with the country's entry into WWI. The year 1917 ushered in modern objects of political concern and an era of rapid change during which such objects began to pass more quickly through the lens of public discourse. It equally marked the decisive ascendancy of today's basic understanding of governance as consisting in foreign policy, on one hand, and a domestic policy centered on the economy, on the other. However, the stream of political discourse that would come to characterize the modern period emerged before the moment of transition to modernity. Although war marked a transition to a new regime of political discourse, the topics of conversation that featured within it were already familiar to contemporaries. Careful observers of late 19th-century and early 20th-century American intellectual history (36, 37) would find confirmation of their readings in these results.

Our modern discourse emerges from a conversation that persists unbroken from the late 19th century into the present. For late 20th-century observers, this stream is recognized as "about" the functions of the liberal-democratic state—concerned with the regulation of business and the financing of public infrastructure. It began, however, in the period centered on 1894, as a moralized discussion

oriented to political and economic reform (Fig. 5). The stream of discussion then split in the following period where the conversation around the regulation of trusts concludes, and another strand oriented to government reform continues eventually becoming a discussion of the welfare state that persists into the present.

The finding is consistent with historical arguments that focus on reformist impulses of the late 19th and early 20th century, as the intellectual, legal, and moral sources of the postwar social order. Progressive ideas did not reshape the political landscape immediately, however; they did so only after the disruption of WWI induced change across multiple domains. The process by which this occurred is beyond this study's scope, but the fact that only under a new regime could the progressive era discourse institutionalize deserves further consideration in the context of other work (36).

More generally, by providing a map of politically relevant categories as they evolved, our study affords a variety of insights into US history. These insights depend on the production of a previously unidentified object: the discursive stream. Although esoteric to academics, political actors and lay observers readily understand political discourse as continuous in this way. Here, we observe that change in salient contents often masks continuity at a higher level. For example, a discussion that at one point in history is about individual rights, the Fourteenth Amendment, and the Supreme Court, is later a conversation about gag rules and development aid, and still later about health insurance and religious employers; we recognize this as the political discourse around abortion. Today, despite a shifting set of key terms, it remains the same thing—a fact revealed by the methods developed and deployed in this article. Although continuity is observable in the SoU, critical discontinuities in political and social discourse are also present: our study reveals a massive transition from 19th century to modern categories of thought as new framings of domestic and foreign policy emerged despite apparently unbroken discussion of fiscal affairs, industry, and state relations.

Moving beyond the SoU, we show that text analysis methods that are oriented to distinguishing analytic levels of interest, and finding ways to capture continuity at these different levels, may provide solutions to classical problems of historical periodization and for understanding social action that traverses the frontiers of historical regimes. Our current digital context is increasingly replete with old documents, textual corpora spanning hundreds of years—in which word use changes for a complex variety of reasons. The network-based text analysis methods we present here can distinguish meaningful from meaningless change in word use, and render higher-level meanings directly interpretable. They are thus uniquely suited to the analysis of corpora that span long historic *durées*.

ACKNOWLEDGMENTS. We thank Christopher Muller, Adam Reich, Shamus Khan, Suresh Naidu, Christopher Blattman, Timothy Shenk, David Hollinger, David Bearman, and the XS workshop at Columbia's Department of Sociology for helpful comments. Maps were produced using the CorText platform. Support from the Interdisciplinary Center for Theory and Empirics at Columbia University is gratefully acknowledged.

1. Tulis J (1987) *The Rhetorical Presidency* (Princeton Univ Press, Princeton).
2. Cohen JE (1995) Presidential rhetoric and the public agenda. *Am J Pol Sci* 39(1):87–107.
3. Hill KQ (1998) The policy agendas of the president and the mass public: A research validation and extension. *Am J Pol Sci* 42(4):1328–1334.
4. Edwards GC, Wood BD (1999) Who influences whom? The President, Congress, and the media. *Am Polit Sci Rev* 93(02):327–344.
5. Zarefsky D (2008) Strategic maneuvering in political argumentation. *Argumentation* 22(3):317–330.
6. Bimes T, Mulroy Q (2004) The rise and decline of presidential populism. *Stud Am Polit Dev* 18(2):136–159.
7. Lazar A, Lazar MM (2004) The discourse of the new world order: Out-casting the double face of threat. *Discourse Soc* 15(2-3):223–242.
8. Teten RL (2003) Evolution of the modern rhetorical presidency: Presidential presentation and development of the State of the Union address. *Pres Stud Q* 33(2): 333–346.
9. Laracey M (2009) The rhetorical presidency today: How does it stand up? *Pres Stud Q* 39(4):908–931.
10. Mohr JW (2013) Graphing the grammar of motives in U.S. national security strategies: Cultural interpretation, automated text analysis and the drama of global politics. *Poetics* 41(6):670–700.
11. Callon M, Courtial JP, Laville F (1991) Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics* 22(1):155–205.
12. Cancho RF, Solé RV (2001) The small world of human language. *Proc R Soc Lond B Biol Sci* 268(1482):2261–2265.
13. Grimmer J, Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Polit Anal* 21(3):267–297.
14. Klingenstein S, Hitchcock T, DeDeo S (2014) The civilizing process in London's Old Bailey. *Proc Natl Acad Sci USA* 111(26):9419–9424.
15. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022.
16. Chuang J, Ramage D, Manning C, Heer J (2012) Interpretation and trust: Designing model-driven visualizations for text analysis. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York), pp 443–452.
17. Newman D, Noh Y, Talley E, Karimi S, Baldwin T (2010) Evaluating topic models for digital libraries. *Proceedings of the 10th Annual Conference on Digital Libraries* (Association for Computing Machinery, New York), pp 215–224.
18. Mimno D, Blei D (2011) Bayesian checking for topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA), pp 227–237.
19. Tang J, Meng Z, Nguyen X, Mei Q, Zhang M (2014) Understanding the limiting factors of topic modeling via posterior contraction analysis. *Proceedings of the 31st International Conference on Machine Learning* (Journal of Machine Learning Research, Microtome Publishing, Brookline, MA), Vol 32.
20. Schmidt BM (2012) Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities* 2(1):49–65.
21. Hofmann T (1999) Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, eds Xing EP, Jebara T (Association for Computing Machinery, New York), pp 50–57.
22. Inouye D, Ravikumar P, Dhillon I (2014) Admixture of Poisson MRFs: A topic model with word dependencies. *Proceedings of the 31st International Conference on Machine Learning* (Journal of Machine Learning Research, Microtome Publishing, Brookline, MA), Vol 32.
23. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (AUAI Press, Arlington, VI), pp 487–494.
24. Wang C, Blei D, Heckerman D (2012) Continuous time dynamic topic models. arXiv: 1206.3298.
25. Wang X, McCallum A (2006) Topics over time: A non-Markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York), pp 424–433.
26. Blei DM, Lafferty JD (2006) Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning* (ACM, New York), pp 113–120.
27. Gao Z, et al. (2011) Tracking and connecting topics via incremental hierarchical dirichlet processes. *Data Mining (ICDM), 2011 IEEE 11th International Conference*, eds Cook D, Pei J, Wang W, Zaïane O, Wu X (IEEE and CPS Conference Publishing Services, Los Alamitos, CA), pp 1056–1061.
28. Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. *Nature* 446(7136):664–667.
29. Hopcroft J, Khan O, Kulis B, Selman B (2004) Tracking evolving communities in large linked networks. *Proc Natl Acad Sci USA* 101(Suppl 1):5249–5253.
30. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci USA* 105(4):1118–1123.
31. Mucha PJ, Richardson T, Macon K, Porter MA, Onnela JP (2010) Community structure in time-dependent, multiscale, and multiplex networks. *Science* 328(5980):876–878.
32. Chavalarias D, Cointet JP (2013) Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PLoS One* 8(2):e54847.
33. Weeds J, Weir D (2005) Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Comput Linguist* 31(4):439–475.
34. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech* 2008(10):10008.
35. Brodsky A (2000) *Grover Cleveland: A Study in Character* (MacMillan, New York).
36. Hollinger DA (1989) *In the American Province: Studies in the History and Historiography of Ideas* (JHU Press, Baltimore).
37. Rodgers DT (1998) *Atlantic crossings* (Harvard Univ Press, Cambridge, MA).
38. Weinstein J (1968) *The Corporate Ideal in the Liberal State, 1900–1918* (Beacon Press, Boston), pp 105–110.
39. Schmid H (1995) Treetagger: A language independent part-of-speech tagger. Institut fr Maschinelle Sprachverarbeitung. *Universitas (Stuttg)* 43:28.
40. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. *Inf Process Manage* 24(5):513–523.
41. Bordag S (2008) A comparison of co-occurrence and similarity measures as simulations of context. *Computational Linguistics and Intelligent Text Processing*, ed Gelbukh A (Springer, Berlin, Heidelberg), Vol 4919, pp 52–63.