MavenRank: Identifying Influential Members of the US Senate Using Lexical Centrality

Anthony Fader
University of Michigan
afader@umich.edu

Dragomir Radev
University of Michigan
radev@umich.edu

Michael H. Crespin
The University of Georgia
crespin@uga.edu

Burt L. Monroe
The Pennsylvania State University
burtmonroe@psu.edu

Kevin M. Quinn
Harvard University
kevin_quinn@harvard.edu

Michael Colaresi
Michigan State University
colaresi@msu.edu

Abstract

We introduce a technique for identifying the most salient participants in a discussion. Our method, MavenRank is based on lexical centrality: a random walk is performed on a graph in which each node is a participant in the discussion and an edge links two participants who use similar rhetoric. As a test, we used MavenRank to identify the most influential members of the US Senate using data from the *US Congressional Record* and used committee ranking to evaluate the output. Our results show that MavenRank scores are largely driven by committee status in most topics, but can capture speaker centrality in topics where speeches are used to indicate ideological position instead of influence legislation.

1 Introduction

In a conversation or debate between a group of people, we can think of two remarks as interacting if they are both comments on the same topic. For example, if one speaker says "taxes should be lowered to help business," while another argues "taxes should be raised to support our schools," the speeches are interacting with each other by describing the same issue. In a debate with many people arguing about many different things, we could imagine a large network of speeches interacting with each other in the same way. If we associate each speech in the network with its speaker, we can try to identify the most important people in the debate based on how central their speeches are in the network.

To describe this type of centrality, we borrow a term from *The Tipping Point* (Gladwell, 2002), in which Gladwell describes a certain type of personality in a social network called a *maven*. A maven is a trusted expert in a specific field who influences other people by passing information and advice. In this paper, our goal is to identify authoritative speakers who control the spread of ideas within a topic. To do this, we introduce MavenRank, which measures the centrality of speeches as nodes in the type of network described in the previous paragraph.

Significant research has been done in the area of identifying central nodes in a network. Various methods exist for measuring centrality, including degree centrality, closeness, betweenness (Freeman, 1977; Newman, 2003), and eigenvector centrality. Eigenvector centrality in particular has been successfully applied to many different types of networks, including hyperlinked web pages (Brin and Page, 1998; Kleinberg, 1998), lexical networks (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Kurland and Lee, 2005; Kurland and Lee, 2006), and semantic networks (Mihalcea et al., 2004). The authors of (Lin and Kan, 2007) extended these methods to include timestamped graphs where nodes are added over time and applied it to multi-document summarization. In (Tong and Faloutsos, 2006), the authors use random walks on a graph as a method for finding a subgraph that best connects some or all of a set of query nodes. In our paper, we introduce a new application of eigenvector centrality for identifying the central speakers in the type of debate or conversation network described above. Our method is based on the one described in (Erkan

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 658–666, Prague, June 2007.

and Radev, 2004) and (Mihalcea and Tarau, 2004), but modified to rank *speakers* instead of documents or sentences.

In our paper, we apply our method to analyze the *US Congressional Record*, which is a verbatim transcript of speeches given in the United States House of Representatives and Senate. The *Record* is a dense corpus of speeches made by a large number of people over a long period of time. Using the transcripts of political speeches adds an extra layer of meaning onto the measure of speaker centrality. The centrality of speakers in Congress can be thought of as a measure of relative importance or influence in the US legislative process. We can also use speaker centrality to analyze committee membership: are the central speakers on a given issue ranking members of a related committee? Is there a type of importance captured through speaker centrality that isn't obvious in the natural committee rankings?

There has been growing interest in using techniques from natural language processing in the area of political science. In (Porter et al., 2005) the authors performed a network analysis of members and committees of the US House of Representatives. They found connections between certain committees and political positions that suggest that committee membership is not determined at random. In (Thomas et al., 2006), the authors use the transcripts of debates from the US Congress to automatically classify speeches as supporting or opposing a given topic by taking advantage of the voting records of the speakers. In (Wang et al., 2005), the authors use a generative model to simultaneously discover groups of voters and topics using the voting records and the text from bills of the US Senate and the United Nations. The authors of (Quinn et al., 2006) introduce a multinomial mixture model to perform unsupervised clustering of Congressional speech documents into topically related categories. We rely on the output of this model to cluster the speeches from the *Record* in order to compare speaker rankings within a topic to related committees.

We take advantage of the natural measures of prestige in Senate committees and use them as a standard for comparison with MavenRank. Our hypothesis is that MavenRank centrality will capture the importance of speakers based on the natural committee rankings and seniority. We can test this claim by clustering speeches into topics and then mapping the topics to related committees. If the hypothesis is correct, then the speaker centrality should be correlated with the natural committee rankings.

There have been other attempts to link floor participation with topics in political science. In (Hall, 1996), the author found that serving on a committee can positively predict participation in Congress, but that seniority was not a good predictor. His measure only looked at six bills in three committees, so his method is by far not as comprehensive as the one that we present here. Our approach with MavenRank differs from previous work by providing a large scale analysis of speaker centrality and bringing natural language processing techniques to the realm of political science.

2 Data

2.1 The US Congressional Speech Corpus

The text used in the experiments is from the United States Congressional Speech corpus (Monroe et al., 2006), which is an XML formatted version of the electronic *United States Congressional Record* from the Library of Congress[1]. The *Congressional Record* is a verbatim transcript of the speeches made in the US House of Representatives and Senate beginning with the 101st Congress in 1998 and includes tens of thousands of speeches per year. In our experiments we focused on the records from the 105th and 106th Senates. The basic unit of the US Congressional Speech corpus is a *record*, which corresponds to a single subsection of the print version of the *Congressional Record* and may contain zero or more speakers. Each paragraph of text within a record is tagged as either speech or non-speech and each paragraph of speech text is tagged with the unique id of the speaker. Figure 1 shows an example record file for the sixth record on July 14th, 1997 in the 105th Senate.

In our experiments we use a smaller unit of analysis called a *speech document* by taking all of the text of a speaker within a single record. The capitalization and punctuation is then removed from the text as in (Monroe et al., 2006) and then the

[1] `http://thomas.loc.gov`

text stemmed using Porter's Snowball II stemmer[2]. Figure 1 shows an example speech document for speaker 15703 (Herb Kohl of Wisconsin) that has been generated from the record in Figure 1.

In addition to speech documents, we also use *speaker documents*. A speaker document is the concatenation of all of a speaker's speech documents within a single session and topic (so a single speaker may have multiple speaker documents across topics). For example within the 105th Senate in topic 1 ("Judicial Nominations"), Senator Kohl has four speech documents, so the speaker document attributed to him within this session and topic would be the text of these four documents treated as a single unit. The order of the concatenation does not matter since we will look at it as a vector of weighted term frequencies (see Section 3.2).

2.2 Topic Clusters

We used the direct output of the 42-topic model of the 105th-108th Senates from (Quinn et al., 2006) to further divide the speech documents into topic clusters. In their paper, they use a model where the probabilities of a document belonging to a certain topic varies smoothly over time and the words within a given document have exactly the same probability of being drawn from a particular topic. These two properties make the model different than standard mixture models (McLachlan and Peel, 2000) and the latent Dirichlet allocation model of (Blei et al., 2003). The model of (Quinn et al., 2006) is most closely related to the model of (Blei and Lafferty, 2006), who present a generalization of the model used by (Quinn et al., 2006). Table 1 lists the 42 topics and their related committees.

The output from the topic model is a $D \times 42$ matrix $\mathbf{Z}$ where D is the number of speech documents and the element z_{dk} represents the probability of the dth speech document being generated by topic k. We clustered the speech documents by assigning a speech document d to the kth cluster where

$$k = \arg\max_{j} z_{dj}.$$

If the maximum value is not unique, we arbitrarily assign d to the lowest numbered cluster where z_{dj} is a maximum. A typical topic cluster contains several hundred speech documents, while some of the larger topic clusters contain several thousand.

2.3 Committee Membership Information

The committee membership information that we used in the experiments is from Stewart and Woon's committee assignment codebook (Stewart and Woon, 2005). This provided us with a roster for each committee and rank and seniority information for each member. In our experiments we use the *rank within party* and *committee seniority* member attributes to test the output of our pipeline. The rank within party attribute orders the members of a committee based on the Resolution that appointed the members with the highest ranking members having the lowest number. The chair and ranking members always receive a rank of 1 within their party. A committee member's committee seniority attribute corresponds to the number of years that the member has served on the given committee.

2.4 Mapping Topics to Committees

In order to test our hypothesis that lexical centrality is correlated with the natural committee rankings, we needed a map from topics to related committees. We based our mapping on Senate Rule XXV,[3] which defines the committees, and the descriptions on committee home pages. Table 1 shows the map, where a topic's related committees are listed in italics below the topic name. Because we are matching short topic names to the complex descriptions given by Rule XXV, the topic-committee map is not one to one or even particularly well defined: some topics are mapped to multiple committees, some topics are not mapped to any committees, and two different topics may be mapped to the same committee. This is not a major problem because even if a one to one map between topics and committees existed, speakers from outside a topic's related committee are free to participate in the topic simply by giving a speech. Therefore there is no way to rank all speakers in a topic using committee information. To test our hypotheses, we focused our attention on topics that have at least one related committee. In Section 4.3 we describe how the MavenRank scores

[2] http://snowball.tartarus.org/algorithms/english/stemmer.html

[3] http://rules.senate.gov/senaterules/rule25.php

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE RECORD SYSTEM "record.dtd">
<RECORD>
  <HEADER>
    <CHAMBER>Senate</CHAMBER>
    <TITLE>NOMINATION OF JOEL KLEIN TO BE ASSISTANT ATTORNEY
    GENERAL IN CHARGE OF THE ANTITRUST DIVISION </TITLE>
    <DATE>19970714</DATE>
  </HEADER>
  <BODY>
    <GRAF>
      <PAGEREF></PAGEREF>
      <SPEAKER>NULL</SPEAKER>
      <NONSPEECH>NOMINATION OF JOEL KLEIN TO BE ASSISTANT
      ATTORNEY GENERAL IN CHARGE OF THE ANTITRUST DIVISION
      (Senate - July 14, 1997)</NONSPEECH>
    </GRAF>
    <GRAF>
      <PAGEREF>S7413</PAGEREF>
      <SPEAKER>15703</SPEAKER>
      <SPEECH> Mr. President, as the ranking Democrat on the
      Antitrust Subcommittee, let me tell you why I support Mr.
      Klein's nomination, why he is a good choice for the job,
      and why we ought to confirm him today.
      </SPEECH>
    </GRAF>
    . . .
    <GRAF>
      <PAGEREF>S7414</PAGEREF>
      <SPEAKER>UNK1</SPEAKER>
      <SPEECH> Without objection, it is so ordered.  </SPEECH>
    </GRAF>
  </BODY>
</RECORD>
```


```
mr presid a the rank democrat on the antitrust subcommitte
let me tell you why i support mr klein nomin why he i a
good choic for the job and why we ought to confirm him
todai
first joel klein i an accomplish lawyer with a distinguish
career he graduat from columbia univers and harvard law
school and clerk for the u court of appeal here in
washington then for justic powel just a importantli he i
the presid choic to head the antitrust divis and i believ
that ani presid democrat or republican i entitl to a strong
presumpt in favor of hi execut branch nomine second joel
klein i a pragmatist not an idealogu hi answer at hi confirm
hear suggest that he i not antibusi a some would claim the
antitrust divis wa in the late 1970 nor anticonsum a some
argu the divis wa dure the 1980 instead he will plot a middl
cours i believ that promot free market fair competit and
consum welfar
the third reason we should confirm joel klein i becaus no on
deserv to linger in thi type of legisl limbo here in congress
we need the input of a confirm head of the antitrust divis
to give u the administr view on a varieti of import polici
matter defens consolid electr deregul and telecommun merger
among other we need someon who can speak with author for the
divis without a cloud hang over hi head
more than that without a confirm leader moral at the
antitrust divis i suffer and given the pace at which the
presid ha nomin and the senat ha confirm appointe if we fail
to approv mr klein it will be at least a year befor we confirm
a replac mayb longer and mayb never so we need to act now we
can't afford to let the antitrust divis continu to drift
final mr presid i have great respect for the senat from south
carolina a well a the senat from nebraska and north dakota
thei have been forc advoc for consum on telecommun matter and
. . .
```

Figure 1: A sample of the text from record 105.sen.19970714.006.xml and the speech document for Senator Herb Kohl of Wisconsin (id 15703) generated from it. The "..." represents omitted text.

1	Judicial Nominations
	Judiciary
2	Law & Crime 1 (Violence / Drugs)
	Judiciary
3	Banking / Finance
	Banking, Housing, and Urban Affairs
4	Armed Forces 1 (Manpower)
	Armed Services
5	Armed Forces 2 (Infrastructure)
	Armed Services
6	Symbolic (Tribute - Living)
7	Symbolic (Congratulations - Sports)
8	Energy
	Energy and Natural Resources
9	Defense (Use of Force)
	Armed Services
	Homeland Security and Governmental Affairs
	Intelligence (Select Committee)
10	Jesse Helms re Debt
11	Environment 1 (Public Lands)
	Energy and Natural Resources
	Agriculture, Nutrition, and Forestry
12	Health 1 (Medical)
	Health, Education, Labor, and Pensions
13	International Affairs (Arms Control)
	Foreign Relations
14	Social Welfare
	Agriculture, Nutrition, and Forestry
	Banking, Housing, and Urban Affairs
	Health, Education, Labor, and Pensions
	Finance
15	Health 2 (Economics - Seniors)
	Health, Education, Labor, and Pensions
	Veterans' Affairs
	Agriculture, Nutrition, and Forestry
	Aging (Special Committee)
	Finance
16	Gordon Smith re Hate Crime
17	Debt / Deficit / Social Security
	Appropriations
	Budget
	Finance
	Aging (Special Committee)
18	Supreme Court / Constitutional
	Judiciary
19	Commercial Infrastructure
	Commerce, Science, and Transportation
20	Symbolic (Remembrance - Military)
21	International Affairs (Diplomacy)
	Foreign Relations
22	Abortion
	Judiciary
	Health, Education, Labor, and Pensions
23	Symbolic (Tribute - Constituent)
24	Agriculture
	Agriculture, Nutrition, and Forestry
25	Intelligence
	Intelligence (Select Committee)
	Homeland Security and Governmental Affairs
26	Health 3 (Economics - General)
	Health, Education, Labor, and Pensions
	Finance
27	Procedural 1 (Housekeeping 1)
28	Procedural 2 (Housekeeping 2)
29	Campaign Finance
	Rules and Administration
30	Law & Crime 2 (Federal)
	Judiciary
31	Child Protection
	Health, Education, Labor, and Pensions
	Agriculture, Nutrition, and Forestry
32	Labor 1 (Workers, esp. Retirement)
	Health, Education, Labor, and Pensions
	Aging (Special Committee)
	Small Business and Entrepreneurship
33	Environment 2 (Regulation)
	Environment and Public Works
	Agriculture, Nutrition, and Forestry
	Energy and Natural Resources
34	Procedural 3 (Legislation 1)
35	Procedural 4 (Legislation 2)
36	Procedural 5 (Housekeeping 3)
37	Procedural 6 (Housekeeping 4)
38	Taxes
	Finance
39	Symbolic (Remembrance - Nonmilitary)
40	Labor 2 (Employment)
	Health, Education, Labor, and Pensions
	Small Business and Entrepreneurship
41	Foreign Trade
	Finance
	Banking, Housing, and Urban Affairs
42	Education
	Health, Education, Labor, and Pensions

Table 1: The numbers and names of the 42 topics from (Quinn et al., 2006) with our mappings to related committees (listed below the topic name, if available).

of speakers who are not members of related committees were taken into account when we measured the rank correlations.

3 MavenRank and Lexical Similarity

The following sections describe MavenRank, a measure of speaker centrality, and tf-idf cosine similarity, which is used to measure the lexical similarity of speeches.

3.1 MavenRank

MavenRank is a graph-based method for finding speaker centrality. It is similar to the methods in (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Kurland and Lee, 2005), which can be used for ranking sentences in extractive summaries and documents in an information retrieval system. Given a collection of speeches $s_1, \ldots, s_N$ and a measure of lexical similarity between pairs $\mathrm{sim}(s_i, s_j) \geq 0$, a similarity graph can be constructed. The nodes of the graph represent the speeches and a weighted similarity edge is placed between pairs that exceed a similarity threshold s_{min}. MavenRank is based on the premise that important speakers will have central speeches in the graph, and that central speeches should be similar to other central speeches. A recursive explanation of this concept is that the score of a speech should be proportional to the scores of its similar neighbors.

Given a speech s in the graph, we can express the recursive definition of its score $p(s)$ as

$$p(s) = \sum_{t \in adj[s]} \frac{p(t)}{wdeg(t)} \tag{1}$$

where $adj[s]$ is the set of all speeches adjacent to s and $wdeg(t) = \sum_{u \in adj[t]} \mathrm{sim}(t, u)$, the weighted degree of t. Equation (1) captures the idea that the MavenRank score of a speech is distributed to its neighbors. We can rewrite this using matrix notation as

$$\mathbf{p} = \mathbf{pB} \tag{2}$$

where $\mathbf{p} = (p(s_1), p(s_2), \ldots, p(s_N))$ and the matrix $\mathbf{B}$ is the row normalized similarity matrix of the graph

$$\mathbf{B}(i, j) = \frac{\mathbf{S}(i, j)}{\sum_k \mathbf{S}(i, k)} \tag{3}$$

where $\mathbf{S}(i, j) = \mathrm{sim}(s_i, s_j)$. Equation (2) shows that the vector of MavenRank scores $\mathbf{p}$ is the left eigenvector of $\mathbf{B}$ with eigenvalue 1.

We can prove that the eigenvector $\mathbf{p}$ exists by using a techinque from (Page et al., 1999). We can treat the matrix $\mathbf{B}$ as a Markov chain describing the transition probabilities of a random walk on the speech similarity graph. The vector $\mathbf{p}$ then represents the stationary distribution of the random walk. It is possible that some parts of the graph are disconnected or that the walk gets trapped in a component. These problems are solved by reserving a small escape probability at each node that represents a chance of jumping to any node in the graph, making the Markov chain irreducible and aperiodic, which guarantees the existence of the eigenvector. Assuming a uniform escape probability for each node on the graph, we can rewrite Equation (2) as

$$\mathbf{p} = \mathbf{p}[d\mathbf{U} + (1 - d)\mathbf{B}] \tag{4}$$

where $\mathbf{U}$ is a square matrix with $\mathbf{U}(i, j) = 1/N$ for all i and j, N is the number of nodes, and d is the escape probability chosen in the interval $[0.1, 0.2]$ (Brin and Page, 1998). Equation (4) is known as *PageRank* (Page et al., 1999) and is used for determining prestige on the web in the Google search engine.

3.2 Lexical Similarity

In our experiments, we used tf-idf cosine similarity to measure lexical similarity between speech documents. We represent each speech document as a vector of term frequencies (or *tf*), which are weighted according to the relative importance of the given term in the cluster. The terms are weighted by their *inverse document frequency* or *idf*. The idf of a term w is given by (Sparck-Jones, 1972)

$$\mathrm{idf}(w) = \log\left(\frac{N}{n_w}\right) \tag{5}$$

where N is the number of documents in the corpus and n_w is the number of documents in the corpus containing the term w. It follows that very common words like "of" or "the" have a very low idf, while the idf values of rare words are higher. In our experiments, we calculated the idf values for each topic using all speech documents across sessions within the

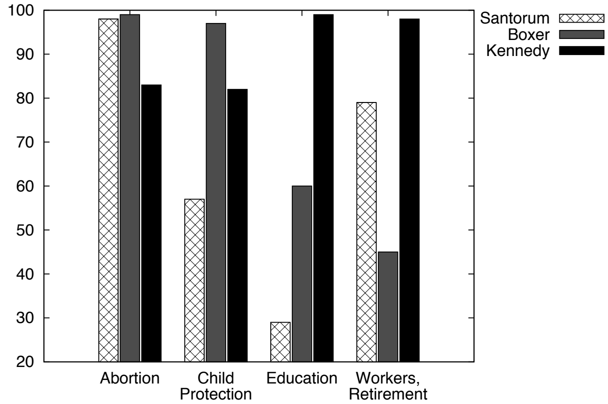

Figure 2: MavenRank percentiles for three speakers over four topics.

given topic. We calculated *topic-specific* idf values because some words may be relatively unimportant in one topic, but important in another. For example, in topic 22 (“Abortion”), the idf of the term “abort” is near 0.20, while in topic 38 (“Taxes”), its idf is near 7.18.

The tf-idf cosine similarity measure $\text{tf-idf-cosine}(u, v)$ is defined as

$$\frac{\sum_{w \in u,v} \mathrm{tf}_u(w)\,\mathrm{tf}_v(w)\,\mathrm{idf}(w)^2}{\sqrt{\sum_{w \in u}(\mathrm{tf}_u(w)\,\mathrm{idf}(w))^2}\sqrt{\sum_{w \in v}(\mathrm{tf}_v(w)\,\mathrm{idf}(w))^2}}, \quad (6)$$

which is the cosine of the angle between the tf-idf vectors.

There are other alternatives to tf-idf cosine similarity. Some other possible similarity measures are document edit distance, the language models from (Kurland and Lee, 2005), or generation probabilities from (Erkan, 2006). For simplicity, we only used tf-idf similarities in our experiments, but any of these measures could be used in this case.

4 Experiments and Results

4.1 Data

We used the topic clusters from the 105th Senate as training data to adjust the parameter s_{min} and observe trends in the data. We did not run experiments to test the effect of different values of s_{min} on MavenRank scores, but our chosen value of 0.25 has shown to give acceptable results in similar experiments (Erkan and Radev, 2004). We used the topic clusters from the 106th Senate as test data. For the speech document networks, there was an average of 351 nodes (speech documents) and 2142 edges per topic. For the speaker document networks, there was an average of 63 nodes (speakers) and 545 edges per topic.

4.2 Experimental Setup

We set up a pipeline using a Perl implementation of tf-idf cosine similarity and MavenRank. We ran MavenRank on the topic clusters and ranked the speakers based on the output. We used two different types granularities of the graphs as input: one where the nodes are speech documents and another where the nodes are speaker documents (see Section 2.1). For the speech document graph, a speaker's score is determined by the sum of the MavenRank scores of the speeches given by that speaker.

4.3 Evaluation Methods

To evaluate our output, we estimate independent ordinary least squares linear regression models of MavenRank centrality for topics with at least one related committee (there are 29 total):

$$\begin{aligned} MavenRank_{ik} = \beta_{0k} + \beta_{sk} Seniority_{ik} + \\ + \beta_{rk} RankingMember_{jk} + \epsilon_{ik} \end{aligned} \quad (7)$$

where i indexes Senators, k indexes topics, $Seniority_{ik}$ is the number of years Senator i has served on the relevant committee for topic k (value zero for those not on a relevant committee) and $RankingMember_{jk}$ has the value of one only for the Chair and ranking minority member of a relevant committee. We are interested primarily in the overall significance of the estimated model (indicating committee effects) and, secondarily, in the specific source of any committee effect in seniority or committee rank.

4.4 Results

Table 2 summarizes the results. “Maven” status on most topics does appear to be driven by committee status, as expected. There are particularly strong effects of seniority and rank in topics tied to the Judiciary, Foreign Relations, and Armed Services committees, as well as legislation-rich areas of domestic policy. Perhaps of greater interest are the topics that do not have committee effects. These are of three distinct types. The first are highly politicized topics for which speeches are intended not to influence

	Topic	$p(F)^a$	$p(\beta_s > 0)^b$	$p(\beta_r > 0)^c$
Seniority and Ranking Status Both Significant				
2	Law & Crime 1 [Violent]	$< .001$	0.016	$< .001$
18	Constitutional	$< .001$	0.003	$< .001$
Seniority Significant				
12	Health 1 [Medical]	$< .001$	$< .001$	0.567
42	Education	$< .001$	$< .001$	0.337
41	Trade	$< .001$	$< .001$	0.087
21	Int'l Affairs [Nonmilitary]	$< .001$	0.007	0.338
9	Defense [Use of Force]	0.002	0.001	0.926
19	Commercial Infrastructure	0.007	0.032	0.332
40	Labor 2 [Employment]	0.029	0.010	0.114
38	Taxes	0.037	0.033	0.895
Ranking Status Significant				
30	Crime 2 [Federal]	$< .001$	0.334	$< .001$
8	Energy	$< .001$	0.145	$< .001$
1	Judicial Nominations	$< .001$	0.668	$< .001$
14	Social Welfare	$< .001$	0.072	0.005
13	Int'l Affairs [Arms]	$< .001$	0.759	0.001
4	Armed Forces 1 [Manpower]	0.007	0.180	0.049

	Topic	$p(F)$	$p(\beta_s > 0)$	$p(\beta_r > 0)$
Seniority and Ranking Status Jointly Significant				
26	Health 3 [Economics]	0.001	0.106	0.064
32	Labor 1 [Workers]	0.007	0.156	0.181
33	Environment 2 [Regulation]	0.007	0.063	0.056
3	Banking / Finance	0.042	0.141	0.579
No Significant Effects of Committee Status				
11	Environment 1 [Public Lands]	0.104	0.102	0.565
22	Abortion	0.419	0.609	0.252
5	Armed Forces 2 [Infrastructure]	0.479	0.267	0.919
24	Agriculture	0.496	0.643	0.425
17	Debt / Social Security	0.502	0.905	0.295
15	Health 2 [Seniors]	0.706	0.502	0.922
25	Intelligence	0.735	0.489	0.834
29	Campaign Finance	0.814	0.748	0.560
31	Child Protection	0.856	0.580	0.718

[a] F-test for joint significance of committee variables.
[b] T-test for significance of committee seniority.
[c] T-test for significance of chair or ranking member status.

Table 2: Significance tests for ordinary least squares (OLS) linear regressions of MavenRank scores (Speech-documents graph) on committee seniority (in years) and ranking status (chair or ranking member), 106th Senate, topic-by-topic. Results for the speaker-documents graph are similar.

legislation as much as indicate an ideological or partisan position, so the mavens are not on particular committees (abortion, children, seniors, the economy). The second are "distributive politics" topics where many Senators speak to defend state or regional interests, so debate is broadly distributed and there are no clear mavens (agriculture, military base closures, public lands). Third are topics where there are not enough speeches for clear results, because most debate occurred after 1999-2000 (post-9/11 intelligence reform, McCain-Feingold campaign finance reform).

Alternative models, using measures of centrality based on the centroid were also examined. Distance to centroid provides broadly similar results as MavenRank, with several marginal significance results reversed in each direction. Cosine similarity with centroid, on the other hand, appears to have no relationship with committee structure.

Figure 2 shows the MavenRank percentiles (using the speech document network) for Senators Rick Santorum, Barbara Boxer, and Edward Kennedy across a few topics in the 106th Senate. These sample scores conform to the expected rankings for these speakers. In this session, Santorum was the sponsor of a bill to ban partial birth abortions and was a spokesman for Social Security reform, which support his high ranking in abortion and workers/retirement. Boxer acted as the lead opposition to Santorum's abortion bill and is known for her support of child abuse laws. Kennedy was ranking member of the Health, Education, Labor, and Pensions committee and the Judiciary committee (which was involved with the abortion bill).

4.5 MavenRank in Other Contexts

MavenRank is a general method for finding central speakers in a discussion and can be applied to areas outside of political science. One potential application would be analyzing blog posts to find "Maven" bloggers by treating blogs as speakers and posts as speeches. Similarly, MavenRank could be used to find central participants in a newsgroup, a forum, or a collection of email conversations.

5 Conclusion

We have presented a technique for identifying lexically central speakers using a graph based method called MavenRank. To test our method for finding central speakers, we analyzed the *Congressional*

Record by creating a map from the clusters of speeches to Senate committees and comparing the natural ranking committee members to the output of MavenRank. We found evidence of a possible relationship between the lexical centrality and committee rank of a speaker by ranking the speeches using MavenRank and computing the rank correlation with the natural ordering of speakers. Some specific committees disagreed with our hypothesis that MavenRank and committee position are correlated, which we propose is because of the non-legislative aspects of those specific committees. The results of our experiment suggest that MavenRank can indeed be used to find central speakers in a corpus of speeches.

We are currently working on applying our methods to the US House of Representatives and other records of parliamentary speech from the United Kingdom and Australia. We have also developed a dynamic version of MavenRank that takes time into account when finding lexical centrality and plan on using it with the various parliamentary records. We are interested in dynamic MavenRank to go further with the idea of tracking how ideas get propagated through a network of debates, including congressional records, blogs, and newsgroups.

Acknowledgments

This paper is based upon work supported by the National Science Foundation under Grant No. 0527513, "DHB: The dynamics of Political Representation and Political Rhetoric". Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

David Blei and John Lafferty. 2006. Dynamic topic models. In *Machine Learning: Proceedings of the Twenty-Third International Conference (ICML)*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Güneş Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.

Gunes Erkan. 2006. Language model-based document clustering using random walks. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 479–486, New York City, USA, June. Association for Computational Linguistics.

L. C. Freeman. 1977. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, March.

Malcolm Gladwell. 2002. *The Tipping Point: How Little Things Can Make a Big Difference*. Back Bay Books, January.

Richard L. Hall. 1996. *Participation in Congress*. Yale University Press.

Jon M. Kleinberg. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677.

Oren Kurland and Lillian Lee. 2005. PageRank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of SIGIR*, pages 306–313.

Oren Kurland and Lillian Lee. 2006. Respect my authority! HITS without hyperlinks, utilizing cluster-based language models. In *Proceedings of SIGIR*, pages 83–90.

Ziheng Lin and Min-Yen Kan. 2007. Timestamped graphs: Evolutionary models of text for multi-document summarization. In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pages 25–32, Rochester, NY, USA. Association for Computational Linguistics.

Geoffrey McLachlan and David Peel. 2000. *Finite Mixture Models*. New York: Wiley.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of the Ninth Conference on Empirical Methods in Natural Language Processing (EMNLP '04)*.

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. 2004. Pagerank on semantic networks, with application to word sense disambiguation. In *Proceedings of the Twentieth International Conference on Computational Linguistics (COLING '04)*, pages 1126–1132.

Burt L. Monroe, Cheryl L. Monroe, Kevin M. Quinn, Dragomir Radev, Michael H. Crespin, Michael P. Colaresi, Anthony Fader, Jacob Balazer, and Steven P. Abney. 2006. United states congressional speech corpus. Department of Political Science, The Pennsylvania State University.

Mark E. J. Newman. 2003. A measure of betweenness centrality based on random walks. Technical Report cond-mat/0309045, Arxiv.org.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66, Stanford Digital Library Technologies Project, Stanford University, November 11,.

Mason A. Porter, Peter J. Mucha, M. E. J. Newman, and Casey M. Warmbrand. 2005. A network analysis of committees in the u.s. house of representatives. *PNAS*, 102(20), May.

Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2006. An automated method of topic-coding legislative speech over time with application to the 105th-108th U.S. senate. In *Midwest Political Science Association Meeting*.

K. Sparck-Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20.

Charles Stewart and Jonathan Woon. 2005. Congressional committee assignments, 103rd to 105th congresses, 1993–1998: Senate, july 12, 2005. `http://web.mit.edu/17.251/www/data_page.html`.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327–335.

Hanghang Tong and Christos Faloutsos. 2006. Center-piece subgraphs: problem definition and fast solutions. In Tina Eliassi-Rad, Lyle H. Ungar, Mark Craven, and Dimitrios Gunopulos, editors, *KDD*, pages 404–413. ACM.

Xuerui Wang, Natasha Mohanty, and Andrew McCallum. 2005. Group and topic discovery from relations and their attributes. In *NIPS*.