The Biggest Cold Warrior: Measuring Influential Nodes by Language Coordination

Eun Seo Jo
Department of History
Stanford University
eunseo@stanford.edu

Abstract

This project introduces a new approach to identifying influential nodes and a method of calculating centrality by taking into account linguistic characteristics of relations between nodes. Using a data set of a correspondence network from a highly stratified group, I test effects of language coordination and simulate the spread of linguistic patterns. This work shows that de jure and de facto influencers may be different in a bureaucratic network. I first calculate edge probabilities based on the likelihood of linguistic pattern adoption then record the size of influence sets. Finally, I compare my results with established centrality measures and find that simulations produce different observations about influential nodes.

1 Introduction

This project is an attempt to integrate understanding about hierarchical language into analysis of influential nodes in a network. NLP literature already supports that lower ranked agents tend to display greater "language coordination" with their superiors.(Danescu-Niculescu-Mizil et al., 2012) Leveraging this information and also corroborating this with my data set, I calculate the level of transference between two nodes, then estimate the global influence level of each node by finding their mean influence sets. Ultimately, this project proposes a new method of calculating centrality of nodes in a network based on the correspondence text between given two nodes. Existing centrality measures such as those of Degree, Eigenvector, and Betweenness only give information about the structure of the network. Diffusion centrality goes further and adds semantic aspects of the networks, taking into account characteristics of the nodes and edges, but this too neglects text information.

While I am using the State Department correspondence corpus as my data set, the aim is to be able to use additional information in non-ranked networks with text to derive hierarchical information and use that to estimate nodes' influence factors in the network. I hope this work spurs further discussion in questions such as: How do diseases or information cascade throughout a network where there is a clear hierarchical structure? How can I compare the actual influence of positions compared to their assigned hierarchical positions? Is personality or position a better indicator of influence?

As a case study, I try to decode which actors in the State Department (such as Secretaries of State and Defense, President, and National Security Adviser) were most influential in the shaping the language of the institution leading up to and during the American intervention in East Asia in the Truman era.

2 Data

The dataset for the project is a collection of diplomatic correspondences within the U.S. State Department from the late nineteenth to the late twentieth century (fig. 1). I have parsed the metadata from a total of 250,000 documents and of these 210,000 are categorized as correspondences, dated and fully labeled with the positions of the sender and recipient. Some of this information had to be imputed by contextual inference. Most of the correspondences are from the high Cold War era, 1950-1970, and each is about 300-800 words in length, discussing topics ranging from trade to military strategizing.

As expected, the Secretary of State ranks first in the volume of correspondences sent ($> 18,000$), followed by the Acting Secretary of State (8,000) and the Department of State (7,000). To mention

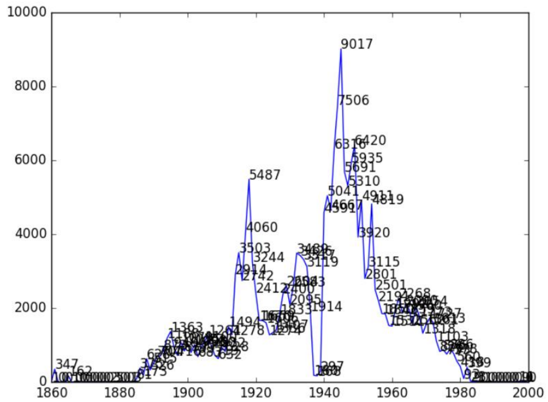


Figure 1: *Graph of document frequency by year*

brief statistics, by centrality measure, the Secretary of State and Acting Secretary of State rank highest again (Eigenvector) but the Assistant Secretary of State for Far Eastern Affairs also shows to be relevant (Eigenvector; Betweenness).

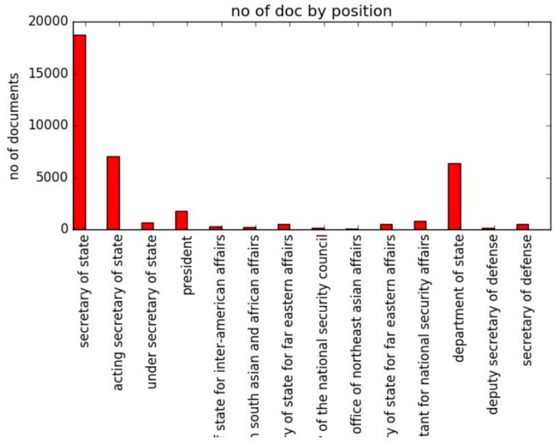


Figure 2: *Out degree by position*

2.1 Structure of the State Department Network

The State Department is a highly bureaucratized and hierarchical institution and provides a rich basis for analysis on structural and linguistic social network patterns.

The ccdf of the correspondence network (fig 3) reveals a strict adherence to a power law distribution with heavy tails ($\alpha = 1.8$). This indicates that while there are few nodes with high degree centrality, meaning few positions with high volume of receiving or sending correspondences, the tail of the distribution is thick such that there is a significant group of elites that have a high volume of communication.

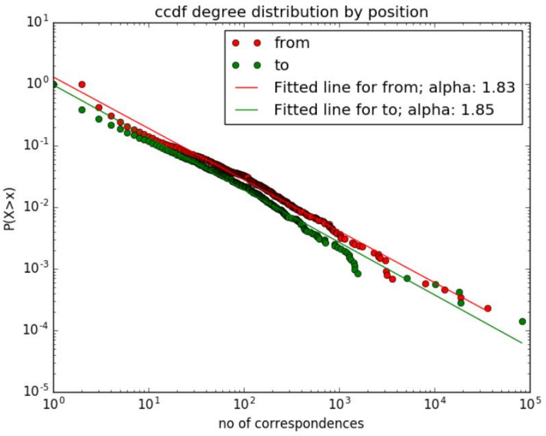


Figure 3: *CCDF degree distribution by position*

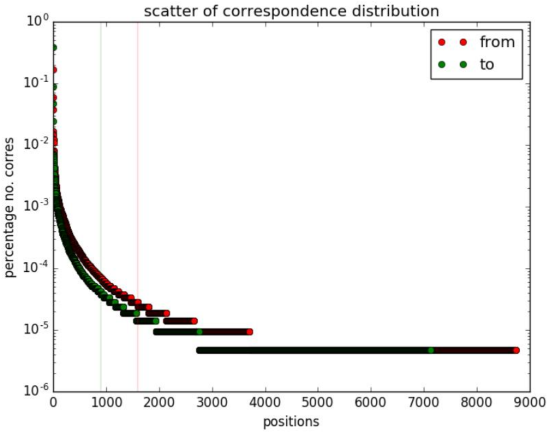


Figure 4: *Scatter plot of correspondence distribution (vertical lines indicate 95% of all correspondences sent and received resp.)*

95% of all correspondences were sent by 1650 most prolific positions while 95% were received by 800 top positions (fig 4). Thus, we can infer from the correspondence network structure that the State Department has a core group of actors who are most prolific and likely to be influential. These actors will be the focus of this study.

2.2 Truman Era

While my future work will be able to test this model on the full data set, for this class project I decided to work with a subset of documents from the 1945-1952 Truman years so that I can assume that the position labels of the correspondence can be taken as the person ID of the correspondence.

The reasoning is that many State Department bureaucrats are appointed at the start of the Presidential term. Without this assumption, I cannot trace the language patterns of nodes because I do not have the name IDs of all positions. I am currently imputing this information on the greater set. In total, 44005 documents are from the Truman era (from January 1945 to January 1953). There are 501 individual posts making up the top 95 of the correspondences sent and 306 individual posts making up the top 95 of the correspondences received. I trimmed the set to include only those sent and received by this group and ended up with 33022 total documents.

3 Literature Review

I engage with two large branches of research in networks - centrality measures and influential identification, these are related but have slightly different primary concerns.

The literature on influential identification tends to be motivated by industry goals of effective injection methods into a network. Many work with social media datasets such as Twitter, the Blogosphere, and Yahoo! Messenger (Bakshy et al., 2011) (Agarwal et al., 2008). While in general, the nodes, which are often service users, are treated as homogeneous entities some works have added more dimensions to analysis by looking at characteristics of posts (length, interestingness, novelty) and the robustness of influence factors over time. There is still no attempt to make strategic use of text attached to the nodes.

One of these works is notable for asking a fundamental question about the meaning of "influence" (Aral et al., 2009). This final paper by Aral et. al. fine-tunes the definition of influence in the spread of contagions by distinguishing homophily-driven diffusion (spread of contagions by utility of commonality) from influence-based diffusion (spread by coercion or inspiration). On their dataset of the spread of a new messenger app, Yahoo! Go, throughout a network of users, Aral et. al find that $> 50\%$ of the contagion spread is explained by homophily and that previous methods tended to overestimate the effect of peer influence.

This work is directly inspired by the questions Aral et. al. address: Are nodes more likely to be infected if they come into contact by a peer influencer (by homophily) or their superior (by influential node)? Given this dataset is that of a strictly hierarchical institution, it is plausible to test whether individuals are influenced differently by the people from different ranks of the bureaucracy.

This work also directly engages with centrality measures that are widely used to estimate influential nodes. Existing measures of centrality such as Degree, Eigenvector, and Betweenness only consider the structure of the network and assume that all nodes and edges have the same characteristics. Diffusion centrality takes into account properties of the nodes and edges and how well these allow diffusion (Kang et al., 2016), but this does not systematically examine language transference between nodes.

Finally, my work bridges NLP and network scholarships by applying NLP knowledge of hierarchical language on networks. Specifically, computational linguists have examined the role of function words, words that entail more stylistic than topical properties, in predicting power relations between two people that have dialogue (Danescu-Niculescu-Mizil et al., 2012). This is a useful indicator of evaluating probability of influence or emulation of one node from another.

4 Methodology

First, I built the network such that a node signifies a member of the State Department by position (eg. ambassador, under secretary of state) and an edge indicates correspondence, where the direction of the edge reflects the direction of the correspondence. This produces a directed, unweighted network of senders and recipients of correspondences.

The overall methodology is first to calculate edge probabilities for every pair of nodes that has correspondence relations then use those probabilities as edge probabilities and the network structure to evaluate each node's centrality measure.

I attempt several ways of calculating probabilities between two nodes – methods of capturing transference of language generally and stylistic linguistic patterns from one node to another.

The first is directly inspired by (Danescu-Niculescu-Mizil et al., 2012) "Coordination measures." Coordination measures quantify how much a certain utterance m in person A triggers an increase in person B's utterance of m:

$$C^m(b,a) = P(\mathcal{E}^m_{u_2 \to u_1} | \mathcal{E}^m_{u_1}) - P(\mathcal{E}^m_{u_2 \to u_1})$$

where u_2 indicates B's utterance and u_1 A's utterance and $\mathcal{E}$ denotes event of the usage of m. Here (Danescu-Niculescu-Mizil et al., 2012) uses marker class m to denote LIWC-derived categories of function words: articles, auxiliary verbs, conjunctions, high-frequency adverbs, impersonal pronouns, personal pronouns, prepositions, and quantifiers. Focus on function words ensures that the speech emulation stems from stylistic similarity, not from the re-iteration of topics. In this work, I specifically address personal pronouns.

Inspired by this model, this work reproduces a more general attempt at capturing linguistic emulation using pointwise mutual information (PMI) produced bi-grams and trigrams.

$$pmi(x,y) = ln\frac{p(x,y)}{p(x)p(y)}$$

I grab the first $n = 100$ results of pmi from the entire corpus of text between given nodes A and B (eliminating all instances of dates, names, and pronouns). Then to reduce the chance of differentiation of word dissemination by topical relevance, I mask out nouns. For every phrase (bigram or trigram) in this corpus, I calculate the following probability from all text $i \to j$:

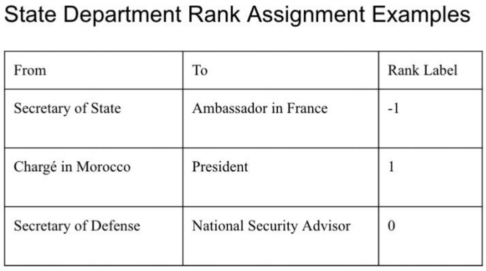
$$P(w_{i \to j}) = \frac{\text{counts of } w_{i \to j}}{N}$$

Where N indicates the sum of all tokens $i \to j$.Then the sum of this probability over every w in the pmi set V:

$$S_{i \to j} = |V|^{-1} \sum_{w \in V} P(w)$$

With these probabilities, I impute the rank difference for every unique relationship $r \in \{-1, 0, 1\}$ where -1 indicates high to low rank correspondence, 1 low to high, and 0 unidentified or neutral. I test for significance of relation between rank difference and language coordination. (See Chart: State Department Rank Assignment Examples) These relative rank measures were done manually based on public, archived information about the State Department organization.

Finally, per node, I simulate at $n = 1000$ to get the average size of influence sets inspired by influence maximization methods in (Domingos et al., 2001). In each round of simulation, I do a simple recursion starting at the root node and infect all neighboring nodes of that node with the given directed edge probabilities. The infected nodes are added to the influence set. I then sum up all influence sets from the simulation and find the average probability of transference.



State Department Rank Assignment Examples

From	To	Rank Label
Secretary of State	Ambassador in France	-1
Chargé in Morocco	President	1
Secretary of Defense	National Security Advisor	0

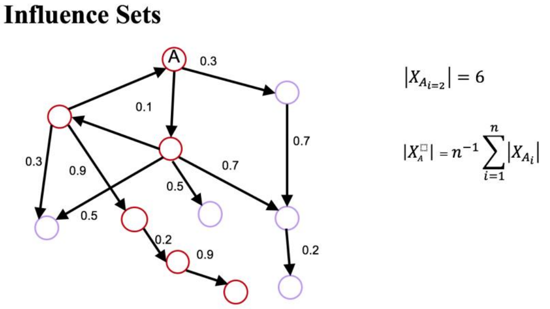


Figure 5: *Influence maximization example*

5 Results

The language coordination centrality measures produce surprisingly different results from existing measures of centrality.

Table 1 shows nodes that ranked the highest by simulation with language coordination edge probabilities, fixed edge probabilities, and random edge probabilities. In comparison with existing centrality measures in table 2, we can see that the simulations produced noticeably different results, particularly with respect to China-related posts. In table 1, I mark with a star, positions associated with China. A third to half of the top ten positions are posts are based in China whereas existing centrality measures show none in the top ten ranked. 5 out of 10 highest ranked nodes by language coordination are based in China.This indicates China-based bureaucrats were more influential in the State Department linguistic sphere. The Truman administration coincides with the Chinese Civil War 1946-1950 and the Chinese Communist Revolution 1949 and thus a generally height-

Language Coordination	**Fixed**	**Random**
Secretary of State	Secretary of State	Secretary of State
Second Secretary of Embassy in China*	Ambassador in the United Kingdom	**Ambassador in China***
Ambassador in Korea	**Ambassador in China***	U.S. Rep at the U.N.
Consul at Taipei*	U.S. rep at the U.N.	Ambassador in France
Ambassador in China*	Ambassador in France	Ambassador in the UK
Acting Secretary of State	**Consul General at Shanghai***	**Consul General at Shanghai***
U.S. High Commissioner for Germany	Acting Secretary State	Acting Secretary of State
Consul General at Shanghai*	**Consul General at Peiping***	**Consul General at Peiping***
Ambassador in Italy	Ambassador in the Soviet Union	Ambassador in India
Consul at Dairen*	Ambassador in India	Ambassador in the Soviet Union

Table 1: *Top 10 nodes ranked by Language Coordination Centrality compared with the results fixed and random simulations*

ened sense of urgency and vigilance surrounding the Cold War in relation to East Asia. Given this historical background, it could be plausible that China-related bureaucrats were more influential in shaping the language of the State Department. It would be worthwhile future work to see if these China-based positions form a structural cluster by community detection and how they relate to the greater network.

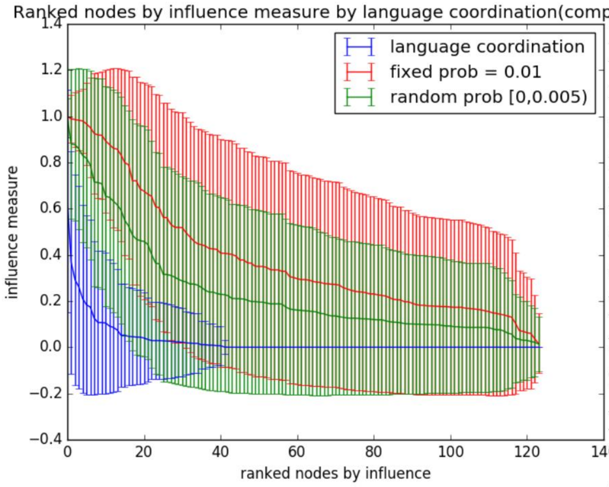


Figure 6: *Ranked nodes by language coordination, fixed, and random measures*

Fig. 6 shows the average influence set size per node (with standard deviation error bars) with edge probabilities calculated by language coordination, with fixed edge probabilities (at 0.01), and random edge probabilities (between [0,0.005)). With language coordination as edge probabilities, most nodes have an average set size of 0 because language coordination probabilities were close to 0.

Language coordination alone, however, did not significantly correlate with pre-registered bureaucratic ranks, showing a p-value of 0.444 in the Mann-Whitney U test of edge probabilities by high-low, low-high rank language transference (high-low indicates transference of patterns from higher ranking to lower ranking nodes).

High to Low	Low to High
0.002	0.001

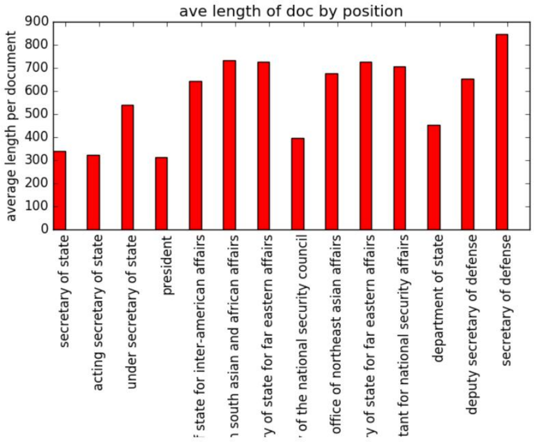
These results could indicate that language transference does not follow the de jure rank order of the bureaucratic institution. In other words, higher ranked by organizational chart may not mean more linguistically influential in this data subset. As another dimension of estimating de facto rank of nodes using linguistic characteristics, I set out to test whether my data set corroborates existing literature on pronoun usage (Kacewicz et al., 2013). (Kacewicz et al., 2013) show that higher ranked members tend to more generously use first-person plural pronouns (”our”,”us”,”we”) than lower ranked members who tend to use first-person singular pronouns (”i”, ”me”, ”my”). In the table below I calculated the probability of the given pronoun usage over all personal pronouns.

pronoun	High rank	Low rank
we	0.210	0.206
us	0.340	0.330
our	0.107	0.109
i	0.147	0.150
me	0.033	0.031
my	0.034	0.036

None of these results are statistically significant at $p < 0.05$. However, in future work, testing this on the entire data set, once cleaned, may yield more interesting results. Furthermore, there are other linguistic indicators of rank difference such as length of correspondence or frequency.

Eigenvector	Betweenness	Degree
Secretary of State	Secretary of State	Secretary of State
Acting Secretary of State	Acting Secretary of State	Acting Secretary of State
Deputy Under Sec of State	Assist Sec of State for Far Eastern Affairs	Under Sec of State
Under Sec of State	Sec of Defense	Deputy Under Sec of State
Assist Sec of State of UN Affairs	US special rep in Europe	Assist Sec of State for Far Eastern Affairs
Assist Sec of State of Far Eastern Affairs	Ambassador in Korea	Assist Sec of State for UN Affairs
Secretary of Defense	Administrator for Economic Cooperation	Sec of Defense
Director of the Policy Planning Staff	Under Sec of State	President
President	Directory of the Policy Planning Staff	Directory of the Policy Planning Staff
Dep Ass Sec of State for Far Eastern Affairs	President	Ambassador at Large

Table 2: *Top 10 nodes ranked by Eigenvector, Betweenness and Degree Centrality*



For future work, the number of correspondences and average length per correspondence are other features that could predict hierarchical rank relations given two nodes. The frequency for instance seems to favor central, Washington-based actors whereas average document length is smaller for the highest ranked positions (Sec of State or President).

6 Evaluation

Works such as this that attempt to produce qualitative, exploratory results about data can be difficult to evaluate as accuracy based results such as prediction are not feasible without ground truth values nor necessarily good indicators of a successful model. However, it is notable that the simulations did not produce any outlandish results far from the existing centrality measures. Many language to rank relations tests did not produce significant test values but trying them again on the entire data set may yield more promising results. Furthermore, the simulations with the language coordination edge probabilities did bring into light interesting historical findings about the overall structure of the State Department and contents of the Truman era. For instance, the positions represented in the top ten of the simulations are generally lower ranked and internationally located (ambassadors) than those that appear in the top ten of the existing measures who are more central and Washington-based. Many of these are China-based, reflecting a particular China-occupied climate in the State Department.(Chang, 2015) As (Chang, 2015) shows China, and American imagination of China, played a large role State Department decision making concerning East Asia in the high Cold War era. In this sense, this centrality measure could be producing more accurate results than do the existing measures.

Overall, adding hierarchical language as a dimension of influence gives historical analysis a richer interpretation of the bureaucracy. This is true especially in a setting where the involved nodes are not homogeneous but have clear hierarchical roles. Finally, I have shown that simulations may actually prove to be better indicators of influential nodes in a correspondence network than centrality measures becaues they allow for edge weights (or frequency of letters) to effect the centrality.

7 Future Work

There are many directions for future work as I continue to organize my data set. One direction is to regress external measures of rank on the results of the model. For instance, the salaries of varying ranks or the year of work experience in the state department can be regressed on the centrality measures with the expectation that higher ranked nodes will tend to also have higher salaries or more experience. Another would be to observe how power structures of the hierarchy over time and results of which words tend to be best the predictors of global rank of a given node.

References

Danescu-Niculescu-Mizil, Cristian, et al. *"Echoes of power: Language effects and power differences in social interaction."*. Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

Danescu-Niculescu-Mizil, Cristian, et al. *"No country for old members: User lifecycle and linguistic change in online communities."*. Proceedings of the 22nd international conference on World Wide Web. ACM, 2013.

Kitsak, Maksim, et al. *"Identification of influential spreaders in complex networks."* Nature physics 6.11 (2010): 888-893.

Aral, Sinan, Lev Muchnik, and Arun Sundararajan. *"Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks."* Proceedings of the National Academy of Sciences 106.51 (2009): 21544-21549.

Bakshy, Eytan, et al. *"Everyone's an influencer: quantifying influence on twitter."* Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu. *"Identifying the influential bloggers in a community."* In Proceedings of the 2008 International Conference on Web Search and Data Mining(WSDM '08). ACM, New York, NY, USA, 207-218. DOI=http://dx.doi.org/10.1145/1341531.1341559

Cha, Meeyoung, et al. *"Measuring User Influence in Twitter: The Million Follower Fallacy."* ICWSM 10.10-17 (2010): 30.

Banerjee, Abhijit, et al. *"The diffusion of microfinance."* Science 341.6144 (2013): 1236498.

Chanhyun Kang et al. *"Diffusion centrality: A paradigm to maximize spread in social networks"* Artificial Intelligence 239 (2016) 7096

Kacewicz, Ewa, et al. *"Pronoun use reflects standings in social hierarchies."* Journal of Language and Social Psychology (2013): 0261927X13502654.

Pedro Domingos and Matt Richardson. 2001. *Mining the network value of customers.* In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '01). ACM, New York, NY, USA, 57-66. DOI=http://dx.doi.org/10.1145/502512.502525

Chang, Gordon H. *Fateful Ties: A History of America's Preoccupation with China.* Cambridge, MA: Harvard UP, 2015. Print.