Large Scale Analysis Final Project

Chen Liang

Project Description

Historically, policy experts in think tanks played an important role in shaping domestic and foreign policies as well as influencing public opinion in the U.S. For example, President Reagan implemented many of the proposals proposed by the Heritage Foundation (Heritage) in its three-thousand-page long publication *Mandate for Leadership*.[i] Policy experts from the Brookings Institution (Brookings), on the other hand, produced reports which blueprinted many major policy decisions including the Marshal Plan[ii] and NASA's space programs[iii].

Despite the profound influence of policy experts in the formation of U.S. politics, academic research on the political influence of policy experts is limited. A particular challenge is the lack of consensus among scholars as to what constitutes a think tank, because policy research institutions differ significantly in their funding measures and targeted audience. Kent Weaver identifies three major types of think tanks.[iv] According to his widely cited theory, "universities without students" such as the American Enterprise Institute (AEI) offer research scholars who either prefer the seclusion of think tanks to teaching responsibilities or dislike the "overwhelmingly liberal"[v] environment in universities a refuge to conduct policy-oriented studies. RAND and the Urban Institute, instead, employ field-specific experts and serve specific needs of policy makers as government contractors. Thirdly, a new generation of policy institutes such as the Heritage Foundation identify themselves as advocacy-oriented and disregard the conventional research principles of neutrality.[vi]

This theory, however, poses two general questions I hope to focus on for my thesis. The first question regards its premise: Can we analyze policy experts from an organizational perspective at all? Does the concept of organizational affiliation really affect the interaction among policy researchers who physically concentrate within only a few blocks in D.C. and hold a wide range of ideological differences *not* always demonstrated in the missions of their institutions? Second, if policy experts in think tanks—at least some of them—still identify themselves as policy researchers, is academic reputation still valued in their social network? How is the academia, i.e. universities, connected to the industry of policy and political ideas, i.e. "universities without students," government contractors, and activists? If not, what expert characteristics determine an expert's influence over its network?

In this project, I try to answer these two questions by proposing a computational way to quantitively analyze the social network of policy experts. Through web-scraping techniques, I construct a Twitter-based follower-following network which can cover about 70% of all policy experts in the three think tanks I chose to start with, i.e. Brookings, AEI, and Heritage. I argue that 1) this network clearly shows the organizational structures of experts in the three think tanks, 2) a comparison of different degree centrality measures implies that, the ability to bridge among other non-friend experts is more important in explaining experts' difference in network influence than their direct connections with other influential experts, and 3) an expert's influence on her inner-organizational network is uncorrelated with academic citation index or tweeting sentiments, but is statistically correlated with the number of followers from government agencies and elected officials. In summary, the results imply that the institutionalization of expert networks transforms the role of policy experts from academic researchers as they were historically to idea brokers[vii] who trade political ideas to targeted audiences without necessarily producing the ideas themselves. However, this conclusion is limited to only three think tanks. Further research will cover the data of more influential think tanks from both ideological sides.

Large Scale Computing Strategies

Web-scraping Twitter

To construct Twitter networks and analyze experts' Twitter usage, I need to obtain a database that contains a full list of Twitter follower and following accounts for each expert and experts' historical tweets. I started by manually matching Twitter users with experts to avoid name duplication and successfully matched about 70% of all experts to their Twitter accounts. For each user, the scraping process was sequential, because I had to extract a cursor from the current page to generate a new URL that points to the next page. This sequential process was time-consuming, and I needed to further manually slow down the scraping process to avoid having my IP blocked by Twitter. Given the situation, I chose to use PyWren with an EC2 instance on Amazon Web Service (AWS) to parallelize the scraping process by scraping multiple users at the same time and then save the data to S3 with boto3. Because we can launch several thousand AWS Lambda functions simultaneously and I only need to scrape about 500 users, I simply mapped each user to a Lambda function and scraped all users at the same time. This parallelization strategy saved me at least 20 hours. In addition, by avoiding the IP blocking program, AWS also significantly increases the sequential scraping process in each function unit.

Sentiment Analysis

To analyze whether experts' sentiments in tweets affect their influence over social network, I used PyWren to scrape at most 200 pages of historical tweets for each expert. I then applied sentiment analysis using the NLTK package and used the multiprocessing—a process-based "threading" interface—with two CPU cores on my own computer calculate the sentiment scores of about 270, 000 tweets. I originally planned to transfer the data to S3 and ran the code on AWS SageMaker notebook which helps improve the training speed of machine learning models. But it turned out that the local parallelization has already been fast enough: I scored all tweets in only about 50 seconds.

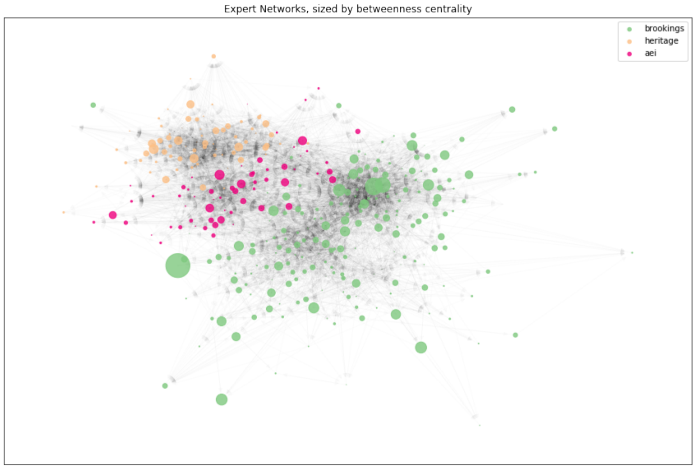
An NLTK sentiment analyzer from Valence Aware Dictionary and sEntiment Reasoner (VADER) applies a lexicon and rule-based model to determine whether a piece of text is positive, negative, or neutral. According to its developers, the model is "specifically attuned to sentiments expressed in social media."[viii] Primarily, it tokenizes a piece of unit text as a bag of words and calculates the percentages of positive and negative words. It generates negative, neutral, and positive scores for each unit text. In most cases, the VADER model properly estimates the level of sentiment in tweets, but with a lack of context, sometimes the model failed severely. For example, in response of the current Black Lives Matter movement, a 1.2-million-likes tweet writes "They'd rather arrest hundreds of American citizens then 3 of their own. Very telling." This is obviously an angry and sarcastic—negative in any means—text, but the model scores it as 83% neutral and only 17% negative because the only negative word is "arrest." As a comparison, the tweet "It is important for advertisers to know what they are sponsoring. Their choices can encourage or discourage bad dangerous propaganda" scores 46% neutral and 37%

negative. The latter tweet is perhaps also negative, but it is obviously not as sentimental as the former tweet. Therefore, further research is needed for better classification.

Results

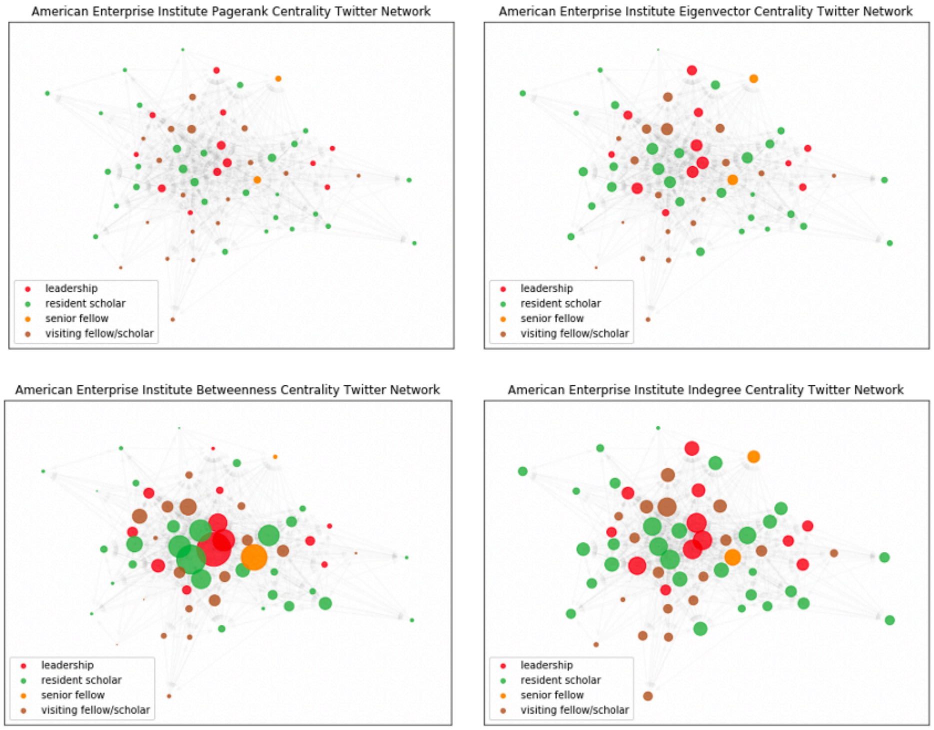
First, I construct an expert network for all experts in the three think tanks and display the result using Spring Layout, which positions nodes using Fruchterman-Reingold force-directed algorithm. As shown in the graph below, expert nodes in each think tank automatically cluster together. We can also observe a spectrum of ideological differences here: The very conservative nodes from Heritage cluster closely at top left, and the relatively diverse—but mostly liberal—nodes from Brookings spread out from the center of the graph to the bottom right.

More rigorous modularity measures would be helpful, but the messages are already promising: First, the Twitter follower-following networks can indeed represent certain inner-organizational network structures. Second, the distribution of betweenness centrality is relatively independent to the inter-organizational positions. That is, the nodes which bridge different think tanks do not tend to be have higher centrality than others. These messages suggest that the social-media based networks can be a nice proxy for the inner-organizational power structure.


Expert Networks, sized by betweenness centrality

Second, I used Gini Coefficient to compare the inequality of different degree centrality measures and took the AEI network as an example. The Gini Coefficients are: 0.34 for Indegree Centrality, 0.096 for Closeness Centrality, 0.38 for Eigenvector Centrality, 0.35 for PageRank Centrality (alpha = 0.85), and 0.66 for Betweenness Centrality. Different measures emphasize different power structures. The inequality level for closeness centrality is extremely low, indicating that nodes are so closely clustered together that for most of the nodes, it does not take a long way to be connected to any other nodes. The inequality levels for indegree, eigenvector, and PageRank centrality are relatively high at about 0.3-0.4, but are not as high as that for betweenness centrality. An over 60 percent Gini coefficient suggests that the most powerful 20 percent nodes control 80 percent of all influence, displaying significant level of influence inequality.

This difference between indegree, eigenvector (PageRank as one of its variations), and betweenness centrality measures imply important patterns in policy experts' social network structures. First, despite some marginal nodes nobody follows—maybe some freshman analysts like research assistants—average research fellows do not tend to actively and comprehensively follow their senior managers, directors, and leadership board members. Second, eigenvector and PageRank centrality measures tend to yield higher unequal results because higher-scored nodes that cluster together tend to reciprocally add to each other's scores. They are not particularly high in our case, probably because senior experts who are supposed to have higher scores do not follow each other closely on the social media. They might be separated based on field specialty, subtle ideological differences, personal relationship, etc. Most importantly, the high level of inequality in betweenness centrality gives implications regarding how to interpret the forms of experts' inner-organizational power. Betweenness centrality measures a node's ability to control information flows: A node with high betweenness centrality plays a more essential role in facilitating the information communications among experts who are not personally friends with each other.


American Enterprise Institute Pagerank Centrality Twitter Network

American Enterprise Institute Eigenvector Centrality Twitter Network

American Enterprise Institute Betweenness Centrality Twitter Network

American Enterprise Institute Indegree Centrality Twitter Network

Finally, I took betweenness centrality as a proxy for experts' influence over their social network and ran regressions to identify the key factors which may explain the inequality in betweenness centrality from an endogenous (non-network) perspective. The independent variables include: the numbers of followers and following accounts controlled by federal government agencies, Congress, news outlets; sentiment in tweets calculated by NLTK; experts' job titles in their affiliated think tanks, Google Scholar citation index, and the frequencies of original and total tweets. In summary, the model explains about 30% total variation in betweenness centrality. Using 0.05 p-value threshold, the only statistically significant factors are the number of followers from Congress (positive), the number of following accounts from government agencies (positive), and experts' job titles (positive). Different think tanks show slightly different results. For Heritage, the model explains over 70% total variation, and betweenness centrality is highly correlated with the number of following accounts from Congress and the frequency of original tweets. For AEI, the model explains only about 23% total variation and the only significant factor is job titles. For Brookings, the model explains about 40% total variation. Those who follow more accounts controlled by government agencies and news outlets and those with higher job titles tend to have higher betweenness centrality.

In summary, I conclude that first, unlike Heritage, AEI and Brookings, the two more established and historical institutions see a higher correlation between institutional arrangements (job titles) and the informal influence over expert networks (betweenness centrality). Second, among the three think tanks, Heritage seems to show a stronger tendency to advocate in Congress through public outlets such as social media, and thus values the experts who follow more Congress members and publish more original contents. Finally, citation index and sentiments in tweets have low predictive power in all models, suggesting that experts who have higher academic reputation or more (or less) emotional tweets do not tend to receive higher influence over expert networks.

[i] Edwards, Lee. *The Power of Ideas*. Ottawa, Illinois: Jameson Books. pp. 41–68. ISBN 0-915463-77-6.

[ii] Nessen, Ron, and Fred Dews. “Brookings's Role in the Marshall Plan.” Brookings. Brookings, April 3, 2018. https://www.brookings.edu/blog/brookings-now/2016/08/24/brookings-role-marshall-plan/.

[iii] Dews, Fred. “Communications, Technology, and Extraterrestrial Life: The Advice Brookings Gave NASA about the Space Program in 1960.” Brookings. Brookings, July 17, 2019. https://www.brookings.edu/blog/brookings-now/2014/05/12/communications-technology-and-extraterrestrial-life-the-advice-brookings-gave-nasa-about-the-space-program-in-1960/.

[iv] Weaver, R. Kent. “The Changing World of Think Tanks: PS: Political Science & Politics.” Cambridge Core. Cambridge University Press, September 2, 2013.

[v] Abelson, D. E. *Capitol Idea: Think Tanks and U.S. Foreign Policy*. Montréal: McGill-Queens University Press, 2014.

[vi] Edwards, Lee. *Leading the Way: the Story of Ed Feulner and the Heritage Foundation*. New York: Crown Forum, 2013.

[vii] Smith, James Allen. *The Idea Brokers: Think Tanks and the Rise of the New Policy Elite*. New York: Free Press, 1993.

[viii] C.J. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, *Eighth International Conference on Weblogs and Social Media* (ICWSM-14). Ann Arbor, MI, June 2014.