Who and What Do Party Activists Talk About?

Steven Morgan
Fangcao Xu
Lulu Peng
Omer Yalcin

May 4, 2018

Abstract

This paper examines the individuals and organizations that party activists are concerned about. While many studies of party activists rely on surveys and interviews of activists such as delegates or party chairmen, this paper utilizes party press and news releases from the 50 state Republican parties. Examining state parties allows for examination of regional variation (i.e. the South versus New England), in addition to variation induced by differences in electoral competition across states (i.e. swing states versus Democratic strongholds like California and New York). We assert that this underutilized data source allows for key insights into what and who party activists are concerned about on a day-to-day basis. Specifically, we examine the entities (both individuals and organizations) named by these texts. We conduct network analysis of a co-occurrence matrix, where entities serve as nodes and releases serve as edges. We find that state GOP parties are concerned with a number of national political actors but are concerned with a litany of state actors as well.

Introduction

Prior work has argued that party activists act as a leading cause of party polarization in American politics. Accounts of contemporary American politics, both in mainstream and academic literature, emphasize the ideological polarization of the two major parties on both economic and social cleavages. Specifically, the Republican Party has moved in a conservative direction while the Democratic Party has moved in a liberal direction on nearly all major policy issues. While there exists a healthy debate regarding mass polarization, work on party and elite polarization converges on less controversial findings: activists and party elites across parties lie on opposite sides of a widened ideological gap (Abramowitz and Saunders 1998; Stonecash, Brewer, and Mariani 2003; Jacobson 2005; McCarty, Poole, and Rosenthal 2006; Theriault 2008; Levendusky 2009; Lee 2009).

Despite this attention to the ideology of party activists, less is known about the relationship between activists and other actors in the party network. Recent work has extended the conceptualization of "party" to include actors not typically considered as embedded in the formal party network. This extended party network includes interest groups, organized groups, the media, donors, and campaign committees (Desmairas, La Raja, and Kowal 2015; Koger, Masket, and Noel 2009; Skinner, Masket, and Dulio 2012). Much of this literature focuses on how candidates navigate this network and how campaign fundraising occurs under this broader conceptualization of party. Less is known how party activists respond to events involving these crucial actors and what signals activists are sending to these actors, as well as to potential candidates and the party-faithful. In this paper, we take a first step at examining how and to whom party activists are addressing. We do this by assuming that party news and press releases are used by party activists to respond to political events. This allows us to utilize this more fine-grained and less-artificial method (compared to survey instruments) to observe how activists respond to (and signal reactions to) actions taken by political entities within the extended party network.

Data

We gather press/news releases from state Republican parties from January 2013 to February 2017 via web scraping. The script we build ignores images, videos, out-links (i.e. to media outlets), and .pdf's. This last data format is potentially concerning, since they contain text. However, the number of .pdf files was minimal and they are only released by a single state GOP party, the Republican Party of New Hampshire. We take two different approaches to retrieving the data. First, we build a web crawler using *Heritrix*, an open-source, highly scalable crawling platform developed by the Internet Archive. However, due to a high level of variation in web page formatting across state party websites, we are unable to effectively produce the desired output (title, date, content).

Instead, we successfully retrieve data via web scraping with *BeautifulSoup* and *Selenium* in Python (for static and dynamic pages, respectively). This produces a corpus of 3,907 press releases, with about 1.3 million tokens, and a vocabulary of just under 400,000 unique terms. However, it should be noted that the data are unbalanced. There exists substantial variation in how often each state Republican party circulates releases. For example, in our time frame, the Pennsylvania GOP produced 1,052 press releases. The GOP of Indiana produced only 8. At this stage, we have not normalized the presence of entities by the number of releases by party or by length of release. On this basis, inferences drawn from our network analysis should be made cautiously. States with many press releases and/or lengthy releases may be driving how often parties are talking about various political elites and organizations. These data contain rich information on political events, partisan taunting, and, important for this work, many references to political figures and organizations.

Methods

Name Entity Recognition

Following the emergence of an unprecedented amount of digital news articles, we need an efficient and accurate way to automatically extract relevant and useful information (e.g. event, time, location, people, news organization etc.) from the big news text dataset. For example, to extract an article's geographic focus, we need to identify the words that are likely linked to geographic locations.

There are a variety of general approaches to identifying named entities in a given text: (1) statistical learning (Zhou and Su, 2001; Burger et al, 2002; Malouf, 2002); (2) using natural language processing (NLP) techniques (Smith and Crane, 2001) to analyze the structure of the text, and (3) scanning the text to search for names listed in a glossary or gazetteer (Gelernter and Mushegian, 2011; Purves et al., 2007; Spitz, Feher and Gertz, 2017) and (4) hybrid approaches (Patrick et al, 2002). The Named-Entity Recognition (NER) (Zhou and Su, 2001) techniques of NLP generally aim to analyze each word (token) of a given sentence using a language specific part-of-speech (POS) tagging process to detect groups of tokens that likely refer to named entities, and seek to classify named entities into pre-defined categories.

The Stanford NER software is used in this paper to extract the people, organization and location from the press releases (in the text format). Java codes have been developed for batch processing all documents and generating a clean format JSON output including key-value pair information below for each press release:

- ID: Index of Original Press Release
- Title: Title of the Press Release
- Time: Date of Publication
- Content: The main body of the Press Release
- Person: People detected in the Press Release
- Organization: Organizations detected in the Press Release
- Geolocation: Place names detected in the content

From the result generated by Stanford NER, there are lots of complete and partial duplicates of the name entities. Some Java functions have been designed in this project to tackle these issues. The duplicate detection enables us to identify and remove those name entities that appeared many times in the same press release (e.g. "Trump" will be deleted when "Donald Trump" exists at the same time).

Geocoding and Hierarchy Detection

Given that the Stanford NER only returns the place names, geocoding technique is used to convert place names (e.g. "1400 Martin St, State College") into geographic coordinates (e.g. 40.7934° N, 77.8600° W), which can be mapped afterwards. The GeoNames and Google API are employed to parse the place name entities. Compared to the GeoNames whose gazetteer is not applicable to street names, the Google Maps API is more robust in detecting fine-grained place names and has an accuracy indicator that tells the scale of the matched object. However, the indicator does not represent the level of confidence in the result, so we could not have a universal score to evaluate the correctness of our geocoding coordinates.

The GeoTxt (Karimzadeh et al, 2013), developed by the GeoVista lab at Penn State Department of Geography, is an open web tool that has integrated the Stanford NLP and Geo-parsing functions and can partly address the ambiguities in the NER. It will be implemented in future research to automatically realize the first two processes (NER and geocoding).
The granularity is also an issue for the place extraction because the results can range from a specific geographical coordinate to a larger region, such as a city or country at various granularity levels for different purposes of analysis. It is thus necessary to develop different strategies to tackle these challenges. In this paper, finding the finest location among multiple locations mentioned in each press release is realized by designing Java functions to parse the components of formatted addresses and detecting the hierarchy based on the smallest unit within the components. That is to say, the higher-level place names will be removed from the geocoding output if there exists smaller grained locations. Taken together, the Java code can automatically take and process all press releases in the folders, and generate a clean JSON output.

Spatial Visualization

The spatiotemporal visualization technique provides a comprehensive dynamic illustration of news coverage over space and time on a map. Recent research has visualized the news events on a map: Gasher (2009) and Howe (2009) both mapped news coverage in and around Phoenix, Arizona, to show how news media portrayed neighborhoods, governmental centers, and ethnic spaces based, in part, on audience preference. Lindgren (2011) mapped local news stories in the Chinese-language newspapers of the Greater Toronto area to show how the ethnic press balanced a 'sense of place' of local and foreign notions. Dörk et al (2008) investigated how coordinated visualizations can enhance search and exploration of news from online RSS feeds by using VisGets for temporal, spatial, and topical (as evidenced by tags) dimensions. In this paper, the R codes and relevant packages (e.g. sp, rgdal, raster, and maptools) are used for mapping all press releases we have collected.

Since the data collection is skewed in the geographic context, many points are distributed along the east coast. Due to the fact that multiple press releases may mention the same location, a point on the map represents a list of documents. It will be difficult to add the labels on the map due to the likelihood of an extreme overlay of information. In future research, we could design an interactive map with functions to segment and recall text information (e.g. ID, Title, Content, Date) at different layers.

Co-occurrence Networks

Co-occurrence networks are utilized to address two questions: a) who and what the press releases concern and b) what the interrelationships among actors tend to be. In such networks, each vertex is the name of a person or organization that appears in at least one press release; an edge between

two vertices exist should they co-occur in the same press release. The weight of each edge corresponds to the frequency of co-occurrence between the two connected vertices.

A few issues in NLP carry over to this step and are addressed with several strategies prior to network construction. First, we filter out press releases in which more than 20 names are mentioned because they tend to influence the interrelationships among actors much more than those with fewer names and would artificially boost the importance of those names characterized by degree and eigenvector centrality. Second, we remove non-alphanumeric characters and leading/trailing Whitespace. Further, by skimming over all the names of "people" captured by NLP, we notice that some words have a low, if any, probability of being a person name (e.g. "Said," "Obamacare," "Took"); thus, we decide to remove them. We also standardize the way in which the most recognized people (e.g. Donald Trump, Barack Obama) and organizations (e.g. Washington Post, New York Times) are called. Another decision has been made to retain only bigrams and trigrams for the people co-occurrence network because most unigrams are unidentifiable and n-grams with $n > 3$ are unlikely to be a person's name.

Network visualization is, for the most part, realized with the package *igraph* in R. Graphs are constructed from edgelists, which is more efficient than the use of adjacency matrices in terms of computation and data storage. Each row in the edgelist is a pair of vertices that co-occur in the corpus of press releases; the tie between them is weighted by the frequency of co-occurrence.

Community detection within graphs can reveal densely connected clusters and highlight the interrelationships among politicians and/or party activists. Five algorithms are used to perform this task, including *walktrap*, *fast greedy*, *spinglass*, *leading eigenvector*, and *infomap*. The results yielded by these algorithms are largely identical with only minor variations. We will interpret the results generated by *walktrap*, which detects subgraphs via random walks.

Results

People Co-occurrence Networks

The complete graph contains 4154 vertices and 23250 edges. With isolates removed, there are 3858 vertices left. Several centrality indices can reveal the importance or impact of each node. The five politicians that rank highest on *degree centrality* are Barack Obama, Rob Gleason, Hillary Clinton, Donald Trump, and Mitt Romney. The rankings on *weighted vertex degree* are slightly different: Tom Wolf ranks the fifth while the other four remain the same. The rankings on *betweenness centrality* is as follows: Barack Obama, Donald Trump, Hillary Clinton, Rob Gleason, and Chris Christie. *Eigenvector centrality* suggests that the most central nodes are Barack Obama, Rob Gleason, Tom Corbett, Pat Toomey, and Hillary Clinton; a similar index, *page rank* shows that Barack Obama, Rob Gleason, Donald Trump, Hillary Clinton, and Tom Wolf are the most prominent actors. As expected, these national (or at least nationally known) figures take central stage in the press releases—these initial analyses present some face validity.

When the vertex size is set to be proportionate to degree centrality, the graph shows a centralized pattern in which a few large vertices occupy the center, surrounded by a large quantity of small ones. When nodes with degree centrality equal or smaller than mean plus one standard deviation are filtered out (resulting in a graph with 138 vertices and 1664 edges), it becomes apparent that the central nodes are nationally known figures including the current and former U.S. presidents and vice presidents, and politicians who ran for president.

The *walktrap* algorithm identifies seven communities in the filtered graph (Figure 1). The six peripheral clusters are politicians from Pennsylvania, Massachusetts, New Hampshire, South Carolina, Maine, and Mississippi, partly because these states are the most active in circulating press releases. The cluster at the center includes national figures as well as party leaders.

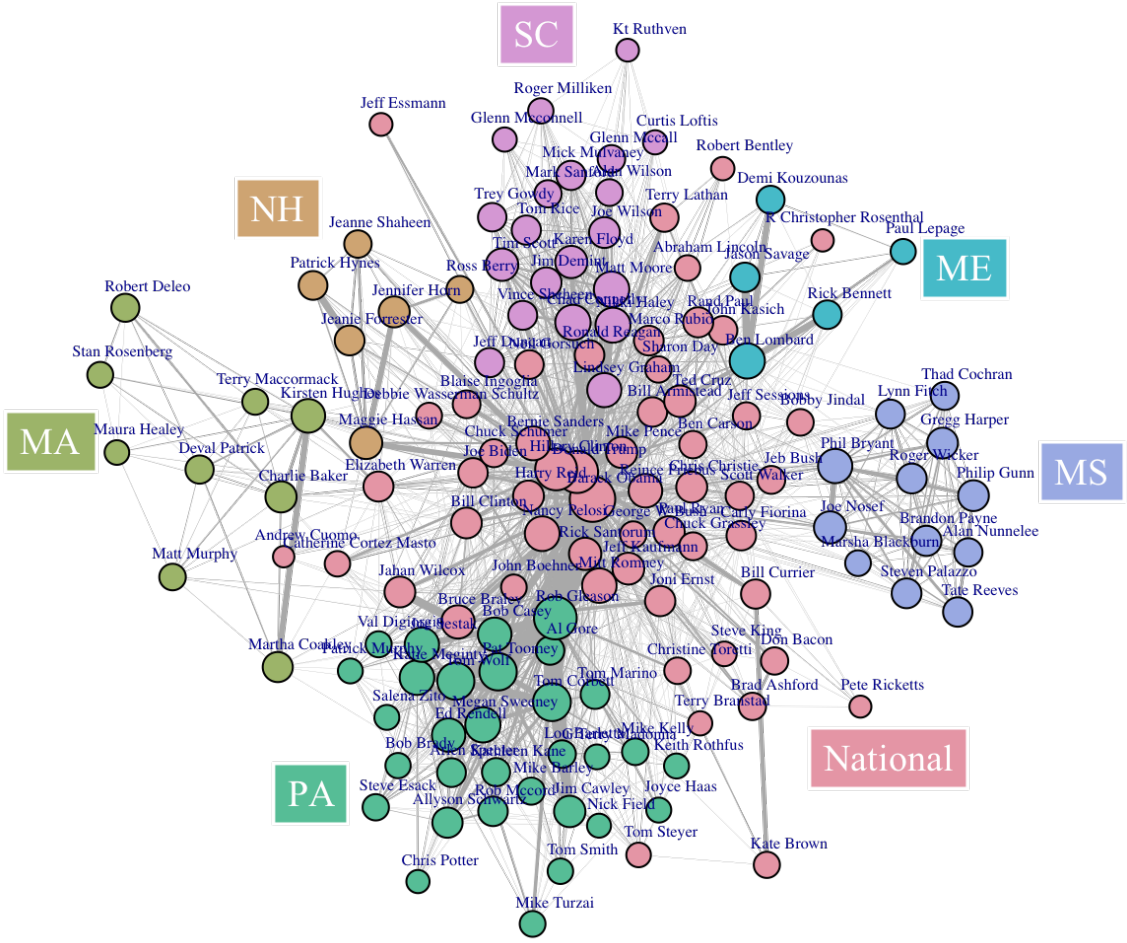

Figure 1. Communities in the People Co-occurrence Network

Organization Co-occurrence Networks

In the organization co-occurrence network we have multiple types of organizations, including government bodies, state and national political parties, media organizations, NGOs and civil society organizations, among others. This means that, unlike "people", who are more or less comparable to one another, organizations differ in their size, purpose, and functions. Nevertheless, the way they interact and engage with one another is worthy of inspection. The graph has 6205 nodes and 54789 edges. When assessed by degree centrality, the most important node is the Senate, followed by the Republican Party of Pennsylvania and the GOP. We can infer that the Republican Party of Pennsylvania is driven by the number of press releases from that state more than anything else. However, most of the top nodes in degree centrality are arguably actually compatible with what would be considered important in reality and include organizations like the House, Congress, Republican National Committee and the Washington Post.

Among the algorithms we run to detect communities, the *walktrap* algorithm's detection seems the most straightforward to interpret, even though others have similar features. The community that has the most "important" nodes, measured in degree centrality, are located in the center of the plot. These include organizations of national importance and therefore oft-mentioned by many state Republican parties. Examples of these are the White House, the House of Representatives, the Senate, the Supreme Court, and FBI, as well as the DNC and RNC. Another cluster is mostly composed of media organizations, such as the Wall Street Journal, the New York Times, the Huffington Post, CNBC, and USA Today. Also recognizable are a number of communities that are identified on the basis of being related to the same state. Some examples of these are the Pennsylvania, Maine, South Carolina, and Mississippi clusters. The formation of these are arguably

partially driven by the high number of press releases available from these states.

Conclusion

All in all, we believe that our project has been a substantial step in using an under-utilized data source – press releases - to understand what party activists talk about. There are a number of ways that we think our project can go further. First of all, as of now, we do not take the temporal dynamic into account in our analysis. However, we know that what state parties talk about does change over time. A good example is "Donald Trump", a central figure in our people network. While we know that Donald Trump has been at the center stage of US politics in the last two years, before 2016 he was much less so. The introduction of a time element would enable us to observe the process of him becoming so central over time, how the various competitors for the Republican nomination for president were talked about by the state Republican parties, which were taken more seriously and which not, and whether states showed differing levels of interest during the nomination process. Another way our study can be extended is its possible use in understanding states that are influential within a national party network. Questions like which state political parties are national conversation starters and which states are the followers might be investigated. Also interesting to see would be how these leading states differ by party and local political dynamics through the inclusion of the Democratic party and measures of local political tendency.

References

Abramowtiz, A.A. and K. L. Saunders, 1998. Ideological realignment in the U.S. electorate. Journal of Politics, 60(3): 634-652.

Desmarais, B. A., La Raja, R. J., & Kowal, M. S. 2015. "The fates of challengers in U.S. house elections: The role of extended party networks in supporting candidates and shaping electoral outcomes: Extended Party Networks in U.S. House Elections." American Journal of Political Science, 59(1), 194-211.

Jacobson, G. C. 2006. Comment on chapter one. In Red and blue nation? Characteristics, causes, and chronology of America's polarized politics, eds. Vol. I, P. Nivola and D. Brady, Washington, D. C., and Stanford, C.A.: Brookings Institution Press and Hoover Institution.

Lee, Frances. E. 2008. Dividers, not uniters: presidential leadership and Senate partisanship, 1981-2004. Journal of Politics, 70: 914-28.

Levendusky, M. S. 2009. The partisan sort. Chicago: University of Chicago Press.

Masket, Seth E. 2009. No Middle Ground: How Informal Party Organizations Control Nominations and Polarize Legislatures. Ann Arbor: University of Michigan

McCarty, N., K.T. Poole, and H. Rosenthal. 2006. Polarized America: the dance of ideology and unequal riches. Cambridge: MIT Press.

Skinner, Richard M,, Seth E. Masket, and David A. Dulio. 2012. "527 Committees and the Political Party Network." American Politics Research Quarterly 40(1): 60-84.

Stonecash, J., M.D. Brewer, and M. Mariani. 2003. Diverging parties: social change, realignment, and party polarization. Boulder, CO: Westview Press.

Theriault, S. 2008. Party polarization in Congress. NY: Cambridge University Press.