Scraping public co-occurrences for statistical network analysis of political elites

Paasha Mahdavi [*]

Abstract

Collecting network information on political elites using conventional methods such as surveys and text records is challenging in authoritarian and/or conflict-ridden states. I introduce a data collection method for elite networks using scraping algorithms to capture public co-appearances at political and social events. Validity checks using existing data show the method effectively replicates interaction-based networks but not networks based on behavioral similarities; in both cases, measurement error remains a concern. Applying the method to Nigeria illustrates that patronage—measured in terms of public connectivity—does not drive national-oil-company appointments. Given that theories of elite behavior aim to understand individual-level interactions, the applicability of data using this technique is well-suited to situations where intrusive data collection is costly or prohibitive.

[*] Assistant Professor, McCourt School of Public Policy, Georgetown University. Address: Old North, Suite 100, 3700 O Street NW, Washington, DC 20057. Telephone: (858) 243-1840. Email: paasha.mahdavi@georgetown.edu. This paper has benefited from discussions with Nicholas Beauchamp, Graeme Blair, Bruce Desmarais, Mark Handcock, John Ishiyama, Franziska Keller, Jeffrey Lewis, Lauren Peritz, Brandon Stewart, Josef Woldense and participants at the 2014 meeting of the Society for Political Methodology, the 2015 meeting of the American Political Science Association, and the UCLA working group on statistical network analysis. All errors are my own.

1 Introduction

How can scholars collect network data on the inner dynamics of political elites in authoritarian and/or conflict-ridden countries? Consider the feasibility of using existing techniques to gather data on elite shuffling in dictatorships (see, for example, Svolik (2012)). Text analysis is challenging given there is no well-defined corpus of texts that can be used to infer ties between elites (other than confidential lists held by political leaders). Applying survey and interview methods may not only be extremely costly but also potentially dangerous to researchers. Enumerators who are surveying elites in Iran, Myanmar, or North Korea are likely to be imprisoned, while interviewers of elites in war-zones such as Afghanistan, Central African Republic, or Syria are at risk of death or severe injury. Researchers engaged in the study of non-democratic and/or war-torn countries—including topics such as regime succession networks, elite patronage networks, military appointments, drugs/arms/human trafficking, and high-level government corruption—will similarly find it difficult to create network data and conduct statistical network analysis.

Political scientists focusing on the developed world are currently equipped with a number of techniques for network data collection on individual actors, ranging from direct observation, to surveys and interviews, to text analysis tools used with archival records. These approaches are well-suited to the study of legislative politics, voter turnout, judicial politics, party politics, and interest group behavior. In particular, current text analysis approaches are ideal in creating network data based on legislative bills, court citations, and party lists. Outside of these areas of study, however, it is difficult to apply these methods.

I provide a new method for gathering network data on political elites in hard-to-reach contexts. Specifically, I build on a measure of social ties—that is, how individuals are connected to one another—that captures the frequency with which individuals attend the same events, organizations, groups, or other social activities. This measure of "co-occurrence" is one possible method to observe what is otherwise a latent characteristic of social connections between individuals. Despite the widespread use of this measure by sociologists dating back to the 1940s, logistical challenges of collecting co-occurrence data on elites in non-democratic or unstable contexts have prevented the adoption of this measure by scholars of comparative politics. To bridge this gap, I combine existing web-scraping tools from computer science with constructs of co-occurrence measures from sociology to propose a technique that assembles network data based on public appearances of individuals at political events covered by reputable online media sources within a given country.

The proposed approach is not without its own limitations, namely measurement error and construct invalidity. Public co-occurrences can imply several types of relationships, only one of which describes a positive, working relationship. They can also capture antagonistic relationships, or "null" relationships with no social interaction despite being present at the same events. Thus it is important to keep in mind, as Grimmer and Stewart (2013, 267) remind us, that "[automated methods] are no substitute for careful thought and close reading and require extensive and problem-specific validation." In the conclusion and in the online appendix, I provide directions for such validation of scraped network data.

2 Techniques for Network Data Collection

2.1 Current Approaches & Limitations

Most quantitative research of network analysis in political science relies on political documents and texts to collect relational data.[1] While techniques for network data collection have yielded fruitful relational datasets thus far, three limitations stand out in using existing approaches to assemble data in non-democracies and war-torn countries.

The first is cost. Obtrusive data collection techniques such as surveys and direct observation can be financially costly to implement for tough-to-reach populations. In authoritarian countries (e.g. Iran, China) and developing democracies (e.g. Russia, Nigeria), it can be prohibitive and potentially dangerous for the researcher to gather data in the field.

[1] A thorough review of existing techniques can be found in Ward et al. (2011) and Lazer (2011).

Second is respondent-induced measurement error. Though measurement error occurs in all social network research, designs that rely on self-reports are particularly vulnerable (Wasserman and Faust, 1994). The validity of self-reported social ties is threatened by respondent bias in whom subjects choose to report as their friends, especially when dealing with political elites.

Archival-based methods solve these problems by capturing behavioral rather than reported ties. Yet these techniques require an identifiable and consistent corpus of text to analyze that limits these techniques to studies of presidents, legislators, party affiliates, and judges. To collect data on patronage appointments in Nigeria, as in the application below, there are no such lists or archival records. The same would be true for trying to collect network data on members of the military regime in Egypt, the extended monarchy network in Saudi Arabia, the clientelism network in Mexico under the PRI, or personnel rotation in dictatorships in general.

2.2 Proposed method

The method for data collection that I propose is based on *measuring social ties as the frequency with which people interact in public.* Ties can be inferred based on co-occurrence, specifically how often given individuals attend the same social events. For political elites, these include activities such as cultural events, fundraisers, campaign banquets, galas, and groundbreaking ceremonies.

While public interactions can provide insight into latent, unobserved characteristics of personal connections between elites, these co-occurrences are valuable in their own right. Public interactions—as opposed to private interactions, which this measure does not capture—can foster deeper personal relationships or they can visibly signal the strength of existing alliances. In 1960s rural Southeast Asia, for instance, public patron-client interactions provided a "physical security mechanism" for clients such that other members of the community could make no mistake about who was under the protection of local chiefs in the absence of formal protective institutions (Scott, 1972, 102–103).

Public co-occurrences are thus informative for studying elite networks in contexts where public knowledge of connectivity is intentional, such as patronage that provides status or physical protection, promotion networks, or hierarchies of power. In contrast, this approach is not useful for studying elite behavior intended to be hidden from the public eye, such as networks of graft, vote-buying, or state structures of repression. The approach is also not well-suited for studying networks of policy processes, such as legislation, judicial outcomes, or coalition formation, where ties tend to form more on the basis of ideological proximity than social closeness.

Co-occurrence networks can be considered a subset of what is often referred to as an "affiliation" network, wherein actors are tied to one another based on their affiliations with the same organizations or events (Wasserman and Faust, 1994, 30–31). In order to avoid the limitations of current techniques of data collection as identified above, a new approach to feasible network data collection is needed. Lee et al. (2010) introduce such a technique using web-based search engines. This relies on the logic that the more often two individuals co-appear in the same news articles and webpages, the more likely that these two individuals interact more frequently when compared to two random counterparts in a given sample.

I build on this approach by applying the technique to collect data on elites in hard-to-reach contexts. Where Lee et al. (2010) measure social ties as co-occurrence in general web pages, I combine the existing affiliation network approach with current text analysis techniques to consider co-occurrence only in the context of physically attending the same events. This is accomplished by using keywords to restrict searches to only capture event attendance as reported online. For example, to capture co-occurrence at political fundraisers in Washington, D.C., I use keywords such as "breakfast fundraiser" or "fundraising dinner" along with domain restrictions to media sites that are known to report on these events, such as `politico.com` or `thehill.com/blogs/in-the-know`.

I construct a sociomatrix—the $n \times n$ matrix where each cell contains the value of a social tie between actors i (rows) and j (columns)—by calculating $x_{ij} = \sum_{g \in \mathcal{G}} c_{ijg}$. Here, x_{ij} represents the value of an undirected tie between i and j; c_{ijg} is a dummy variable for whether a given webpage g contains both i and j (and any additional keywords); and $\mathcal{G}$ is the set of all webpages in a given search. The diagonals of the sociomatrix are given by $x_{ii} = x_{jj}$, which is simply the number of webpages $g \in \mathcal{G}$ that contain an individual i's name. Each tie x_{ij} is the count of webpages satisfying the keyword criteria that contains the names of both i and j. When using Google, this is referred to as the number of "hits" for a given search term.

To fill the sociomatrix, I create an algorithm to iteratively search over all $\binom{n}{2}$ possible undirected pairs, written in `perl` and to be integrated into an `R` package for ease of use. The algorithm is as follows:

1. Create list of individuals in network population (n)
2. Specify search criteria
3. Iteratively search pairs of individuals (i, j)
4. Record number of unique articles paired individuals appear in together ($\sum_{g \in \mathcal{G}} c_{ijg}$)
5. Randomly sample individual page results and calibrate search keywords accordingly
6. Repeat 2-5 until randomly sampled pages are appropriate as desired

Importantly, the search criteria in step 2 are used to capture individuals appearing in relevant events and reduce repetition of media stories by restricting site domains. Step 5 is critical to ensuring that the search criteria are appropriate, similar to the procedure in human-assisted text analysis algorithms. Here, the researcher combs through randomly sampled pages to determine if the resulting pages capture co-occurrence at events. I provide code and full details on these steps in the appendix, with attention as well to conducting sensitivity analyses on resulting networks scraped across different search criteria.

2.3 Assessing conceptual accuracy

Like any method of network data collection, the proposed approach is subject to measurement error. First, measuring ties through co-occurrence may not be accurate in assessing the "true" direction of social ties between individuals. For example, political opponents may frequently attend the same events but never interact with one another, yet a measure of co-occurrence would suggest a strong social tie between these individuals.

Second, despite iteratively refining keywords in the algorithm, it is possible that the searches are still resulting in irrelevant webpages or media stories that list both individuals i and j but in separate parts of the text (e.g. multiple different articles in the same webpage).

Third, co-occurrence as captured by media reports may be subject to reporting bias: the media may be over-reporting the attendance of certain "celebrity" elites while under-reporting the presence of less popular elites, attenuating the connectivity measure of individuals in the latter group towards zero. Validating networks scraped across as many source sites as possible may reduce this bias to some extent. Still, co-occurrence measures cannot differentiate these individuals from well-connected elites who, for whatever reason, do not attend social events. Users should thus exercise caution in interpreting singletons (nodes with zero edges) in contexts where this is likely to occur, such as measuring connectivity of female elites in networks of strongly patriarchal societies where women are discouraged from attending social events.

For these reasons, it is necessary to compare the network data output from the proposed method to existing network data as collected by one of the conventional approaches identified above. I validate the method's conceptual accuracy with five existing network datasets: *directly* using the North Korean guidance visit network (Ishiyama, 2014) and US Senate press event network (Desmarais et al., 2015), and *indirectly* using US Senate co-sponsorships (Fowler, 2006), US House caucus memberships (Victor and Ringe, 2009), and Mexican board memberships (Avina-Vazquez and Uddin, 2013). I provide a full description and results from these validation exercises in the appendix.

In short, these comparisons indicate that the proposed method is not appropriate for US Congressional co-sponsorship and caucus membership network data but is accurate in creating network data based on co-occurrences at events as in North Korea, in Mexican board meetings, and to a lesser extent at US Senate press events. This is not to say that there is no measurement error in the approach in these latter cases. Refining the algorithm and cross-validating with existing data where available is crucial to improving conceptual accuracy. The burden is thus on the researcher to assess the validity of the algorithm's output based on the characteristics of the population of interest.

3 Patronage appointments in Nigeria

I now apply the proposed tool to address the question: Do leaders in dictatorships and developing democracies use government appointments as a tool to dole out patronage (Bueno de Mesquita et al., 2003)? I test an observable implication of this question by looking at the appointment process to some of the most lucrative government positions: board membership in national oil companies. Specifically, I test whether appointments to lucrative state-owned enterprise positions are based on one type of patronage-based linkage: political connections *via* social connectivity in public co-appearances.

I choose Nigeria as a testing ground given it is an extreme case of oil-related patronage. If social connections are not linked to government appointments in Nigeria, this would challenge a long-standing narrative that the country's national oil company (NNPC) exists solely as a money pit for the president and his hand-picked cronies (for a review, see Victor et al., 2012, 701–52). The Nigerian case provides a convenient application for board appointments given that appointments are made concurrently—and given the preponderance of paparazzi-style reporters, Nigerian newspapers frequently report on the attendance of elites at major social/political events. The high profile of public events thus allows for a more precise estimation of social ties using the proposed data collection method.

In July 2012, president Goodluck Jonathan made eight appointments to the NNPC Board. Based on previous appointments, the population of potential appointees consists of all 31 executive cabinet ministers and 20 NNPC executives, making for a network population of $n = 51$. I apply the same strategy to the November 2015 appointments by president Mohammadu Buhari. For the 2015 network, I include all 31 executive cabinet ministers (none of whom served in Jonathan's cabinet) and 20 NNPC executives, including prior board members.

Because no existing network data on the Nigerian oil elite has been collected to date, I apply the proposed algorithm to create relational data based on the list of 51 possible appointees for 2012 and 2015, respectively. Social ties are measured based on pair-wise searches with the following search restrictions:

- Dates: one year prior to the board appointments announcement (1 July 2011 – 31 June 2012 and 26 June 2014 – 25 June 2015);
- Keywords: "fundraising dinner", "groundbreaking ceremony", "gala", "banquet", "campaign event";
- Newspaper restrictions: ngrguardiannews.com, punchng.com, and vanguardngr.com.

Based on informal interviews with Nigerian oil experts,[2] these three newspapers provide near-comprehensive accounts of social political events. Events identified via the search algorithm are likely representative of media-covered events in the country. Events not captured are thus not "high-profile enough" for coverage—implying that these are not necessary to estimate co-occurrence at salient social political events.

Yet this remains a key limitation of the method. The sample of media-covered reports searched by the algorithm is very much at the discretion of the researcher. One can favor an approach akin to random sampling by choosing more restrictive domain terms, or on the other end of the spectrum, one can choose fewer restrictions to generate a more census-like

[2] Interviews conducted via email in October 2013 with four anonymous oil consultants based in Nigeria.

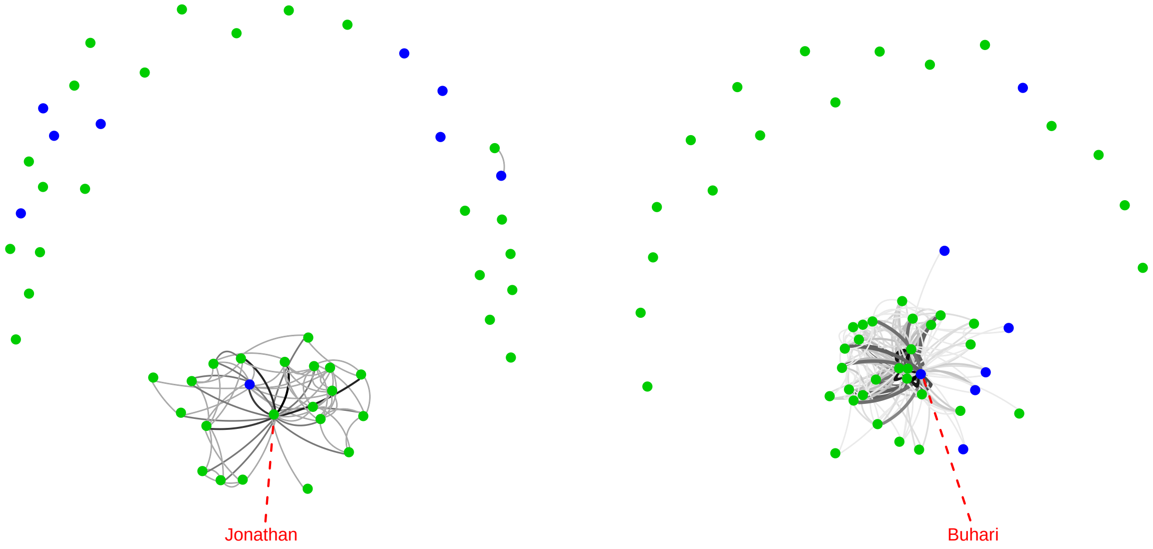

Figure 1: Network graph for potential appointees to NNPC board in 2012 (left) and 2015 (right). Appointees are shown in blue. Edges are weighted by frequency of co-occurrence (darker curves = greater co-occurrence).

sample. Here, I straddle between these extremes, but closer to the latter approach given the high coverage of these three Nigerian newspapers.

In Figure 1, I plot the resulting networks with pairwise edges weighted by co-occurrences. The graphs illustrate that board appointees are not central in either network. In 2012 only one appointee—Diezani Alison-Madueke, managing director of NNPC and the board—was socially connected to the president and other individuals prior to her appointment. The rest are singletons, with no connections to other individuals. In 2015 five appointees co-occurred with other individuals, though they are clearly peripheral compared to rest of the network. Comparing the number and strength of ties of appointees to non-appointees illustrates this point further. In 2015, non-appointees averaged 27 co-occurrences compared with appointees averaging only 4 co-occurrences.

Interestingly, the 2015 network is denser than the 2012 network—not only with fewer singletons but also with stronger ties within the main cluster. In Buhari's administration, public co-occurrences happen with far greater frequency than in Jonathan's administration. This pattern perhaps reflects the former's stronger social ties with his cabinet members given

Buhari's previous leadership during Nigeria's military dictatorship of 1983–1993.

A more rigorous approach to testing this hypothesis is to apply the Exponential-family Random Graph Model (ERGM) to the network data (Cranmer and Desmarais, 2011). Here I apply a somewhat "roundabout" method, which I describe in more detail in the appendix: in brief, I model edges using co-occurrences (since this is a dyadic variable) and board appointments as a covariate (since this is a nodal variable). Results from different specifications (Appendix Table 2) support the descriptive findings: board appointments are not positively correlated with social connectivity. In two of these models, there is evidence for a *negative* relationship, suggesting that appointees are less connected than non-appointees.[3] A more conventional approach using logit models—where board appointment is the outcome and network connectivity is the explanatory variable—shows similar results (Appendix Table 3).

While the research design is not strong enough to make causal inferences, these results suggest that being appointed to the NNPC board of directors is not associated with one's publicly-visible political connectedness to others in the network. There is little evidence of the president making nepotistic appointments to the government's lucrative oil company, in contrast to previous qualitative scholarly work on NNPC (cf. Gillies, 2009). Results from network analysis instead suggest that patronage appointments to NNPC may either be exaggerated, or determined by factors other than public connectivity, such as private ties with loyalists often outside the public light.

4 Conclusion

The technique proposed here provides new opportunities for collecting network data in hard-to-reach contexts. This ranges from populations in developed countries for which existing approaches are costly or infeasible to network populations in conflict-ridden and/or authoritarian countries where on-the-ground research is prohibitive. I have provided one application

[3] To account for the possibility of Google's increasing coverage of reports over time, the number of individual results ("self-hits") is added as a control to these models.

of the method in the context of patronage networks within Nigerian government appointments to state-owned enterprises. In a separate study, I apply the same approach to measure dynamics of the inner elite in North Korea (Mahdavi and Ishiyama, 2016).

The method is conceptually accurate as cross-validated with existing network datasets on public social and political interactions—as in the case of Kim Jong-un's guidance visit network—yet the proposed method can still suffer from measurement and sampling errors. The search algorithm by default is limited to sources that are published in searchable online web pages.[4] Information on event co-occurrence that is published offline is thus omitted. Restricting domain names during iterative searches can also make the resulting sample unrepresentative of the underlying population. Ultimately, the algorithm will produce either a full census of results, a random but representative sample of results, or a random, non-representative sample. Future research can provide improvements to reduce this kind of sampling error.

Second, there is the larger question of whether co-occurrence at events is an appropriate measure of social ties. How often individuals co-attend the same events may not necessarily be indicative of how closely individuals are tied in terms of friendship, ideology, or collaboration, especially if these ties are fostered in private meetings or settings. Rivals may attend the same events, for instance, and thus more co-occurrences can imply a social tie that is not based on ideological or personal closeness. Indeed, all measures used in social network analysis suffer from this conceptual problem given that ties are a latent, unobservable characteristic. More and more cross-validation of the proposed method with existing network data with different measures of social ties can help address this concern.

Nonetheless, co-appearances inherently mean different things in different contexts. Co-occurrences at inspection visits in autocracies can be symbolic of deep ties to the dictator, while in a democracy co-appearances at such ceremonies can represent endorsements of projects by political leaders. Popularity may incentivize co-appearances in social events by

[4] The technique can be applied without relying on the Google API, such as parsing co-occurrences in online news sources directly or indirectly via media databases.

local politicians to “see and be seen” with politically popular leaders. Or they may create a negative stigma, such that aspiring politicians may want to appear as outsiders who are not invited to “clubby” events, or may want to remain inconspicuous in public to avoid arousing suspicion of illicit activities. Co-appearances may also simply reflect routines in daily political life, such as attending open meetings, public hearings, and press events. Given the range in how scholars conceptualize social co-appearances, the method I have proposed here will be applicable to a wide number of settings in comparative politics.

References

Avina-Vazquez, C. R. and S. Uddin (2013). Network of Board of Directors in Mexican Corporations: A Social Network Analysis.

Bueno de Mesquita, B., A. Smith, R. Siverson, and J. Morrow (2003). *The Logic of Political Survival*. Cambridge: MIT Press.

Cranmer, S. J. and B. A. Desmarais (2011). Inferential Network Analysis with Exponential Random Graph Models. *Political Analysis 19*(1), 66–86.

Desmarais, B. A., V. G. Moscardelli, B. F. Schaffner, and M. S. Kowal (2015). Measuring legislative collaboration: The senate press events network. *Social Networks 40*, 43–54.

Fowler, J. H. (2006). Connecting the Congress: A Study of Cosponsorship Networks. *Political Analysis 14*(4), pp. 456–487.

Gillies, A. (2009, February). Reforming corruption out of Nigerian oil? *Chr. Michelson Institute U4 Brief 2*.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis 21*, 267–297.

Ishiyama, J. (2014). Assessing the leadership transition in north korea: Using network analysis of field inspections, 1997-2012. *Communist and Post-Communist Studies 47*, 137–146.

Lazer, D. (2011). Networks in political science: Back to the future. *PS: Political Science & Politics 44*(01), 61–68.

Lee, S. H., P.-J. Kim, Y.-Y. Ahn, and H. Jeong (2010, July). Googling Social Interactions: Web Search Engine Based Social Network Construction. *PLoS ONE 5*(7), e11233.

Mahdavi, P. and J. Ishiyama (2016). Dynamics of the inner elite in dictatorships: Evidence from north korea.

Scott, J. C. (1972). Patron-client politics and political change in southeast asia. *American Political Science Review 66*(1), 91–113.

Svolik, M. (2012). *The Politics of Authoritarian Rule*. New York, N.Y.: Cambridge University Press.

Victor, D. G., D. Hults, and M. C. Thurber (Eds.) (2012). *Oil and Governance: State-owned Enterprises and the World Energy Supply*. Cambridge University Press.

Victor, J. N. and N. Ringe (2009). The Social Utility of Informal Institutions Caucuses as Networks in the 110th US House of Representatives. *American Politics Research 37*(5), 742–766.

Ward, M. D., K. Stovel, and A. Sacks (2011). Network analysis and political science. *Annual Review of Political Science 14*, 245–264.

Wasserman, S. and K. Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.