# Online Appendix for "Scraping public co-occurrences for statistical network analysis of political elites"

July 26, 2017

## Assessing conceptual validity with existing network data

If measurement error using the proposed algorithm is high (for reasons specified in the main text), then the network data resulting from Google searches should not conform with existing network data. The same would be true if the co-occurrence network is measuring an inherently different social network. For this reason, I validate the proposed method using existing network data from five different datasets: Ishiyama (2014), Desmarais et al. (2015), Fowler (2006), Victor and Ringe (2009), and Avina-Vazquez and Uddin (2013).

Ishiyama (2014) manually codes data on the co-appearances of North Korean elites at "guidance visits" or "inspections" led by Kim Jong Un in 2012, as documented by the state-run KCNA. Those elites appearing more frequently with the Great Leader are hypothesized to be more central within Kim's inner circle of elites. The 2012 network consists of 41 actors (including the leader), giving a $41 \times 41$ sociomatrix of ties. I compare this matrix to one created using the proposed scraping method on the same list of 41 elites. The algorithm scraped only the South Korean newspaper *North Korea Leadership Watch* website (which mirrors reports published by the KCNA), included the keywords `visit+and+(inspection+OR+guidance)`, and was restricted to the year 2012. I set these specific restrictions to parallel the method of data collection in Ishiyama (2014), whose network comprises individuals co-appearing with Kim Jong Un in inspection visits or guidance visits during his first year in power using KCNA daily reports. As such, there are no keyword refinements as with other cases below.

Figure 1 shows the correlation between the two networks of co-occurrences, with each point representing the valued-edge of each network dyad. I log both axes given the slight skew towards officials nearly always appearing with the great leader.[1]

In Figure 2, I focus on the elites with fewer than 9 co-occurrences with Kim Jong Un according to the Ishiyama dataset. Here I plot the dyad values in their raw units (not logged) along with the median co-occurrence from the proposed method with 95% confidence bands. Both plots indicate the validity of the proposed method compared to the existing network data. Pairs with no co-occurrences in the Ishiyama data largely have no co-occurrences in the scraped data; those with more than one co-appearance with Kim Jong Un at an inspection visit largely correspond to more than one co-occurrence in the scraped data. A log-log regression of the scraped measure and co-occurrences from the Ishiyama dataset gives a coefficient of 0.93 (SE = 0.03, $t = 34.65$) suggests a nearly one-to-one relationship. A linear regression without logging either variable gives a coefficient of 1.83 (SE = 0.06, $t = 29.05$), indicating that for every additional co-appearance in the Ishiyama data, there is an expected 1.83 additional co-occurrences in events scraped from the web.

Still, the plots show much larger variance in the scraped co-occurrence measure compared to Ishiyama's measure. Part of this variation can be attributed to measurement error issues discussed in the main text, with the remaining variance likely coming from differences between what is reported by *North Korea Leadership Watch* and the KCNA. This case, as with the others discussed here, illustrates the need for researchers using the scraped co-

[1] This list includes Ri Jae Il, Jang Song Thaek, and Choe Ryong Hae. The latter two have since been ousted from the inner circle, with reports that Jang Song Thaek (the great leader's uncle) was executed in December 2013. See https://www.yahoo.com/news/nkorea-executes-kim-jong-un-39-uncle-214524268.html?ref=gs.

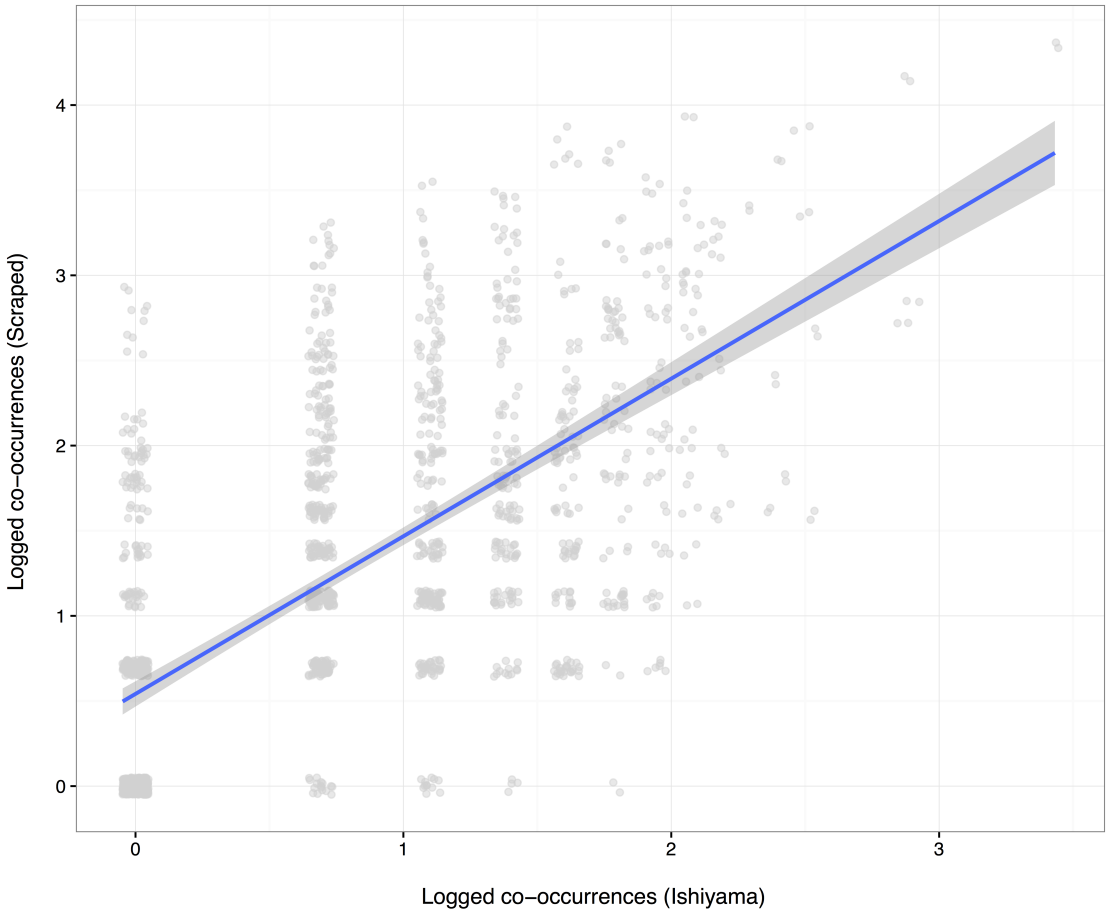


*Figure 1: Validating the proposed method with the Ishiyama (2014) North Korean elite network. A linear fit with a 95% confidence band is shown to compare co-occurrences from the Ishiyama data (x-axis, logged) to co-occurrences from the proposed method (y-axis, logged). The sample includes all possible combinations of 41 elites, with a total effective sample of 820 undirected dyads, represented in the plot by gray points. Points are jittered to help alleviate overplotting.*

occurrences technique to assess competing reasons for possible measurement error based on their substantive knowledge of the context at hand.

Desmarais et al. (2015) captures public co-appearances by senators at press events; these appearances are conceptually similar to co-occurrences at social events. Here I compare the full sociomatrix of the senate press event network from the 110th Congress (2007–2009) to a sociomatrix assembled using the proposed method on the same list of 102 senators.[2]

The search algorithm was restricted to the daterange for the 110th Congress and included the keywords `gala+OR+ceremony+OR+fundraiser+OR+groundbreaking+event` after several runs and refinements.[3] Because of the subjectivity of choosing keywords—or in other cases,

[2]John Barrasso (R-WY) was appointed in June 2007 after the death of Craig Thomas; Roger Wicker (R-MS) was appointed in December 2007 after the resignation of Trent Lott.

[3]These refinements include dropping "fundraiser", replacing "groundbreaking event" with either "press event" or any event in general ("event"), only searching for "press event", and searching for "dinner" instead of "ceremony". The list of keyword variations is presented in the caption for Table 1.

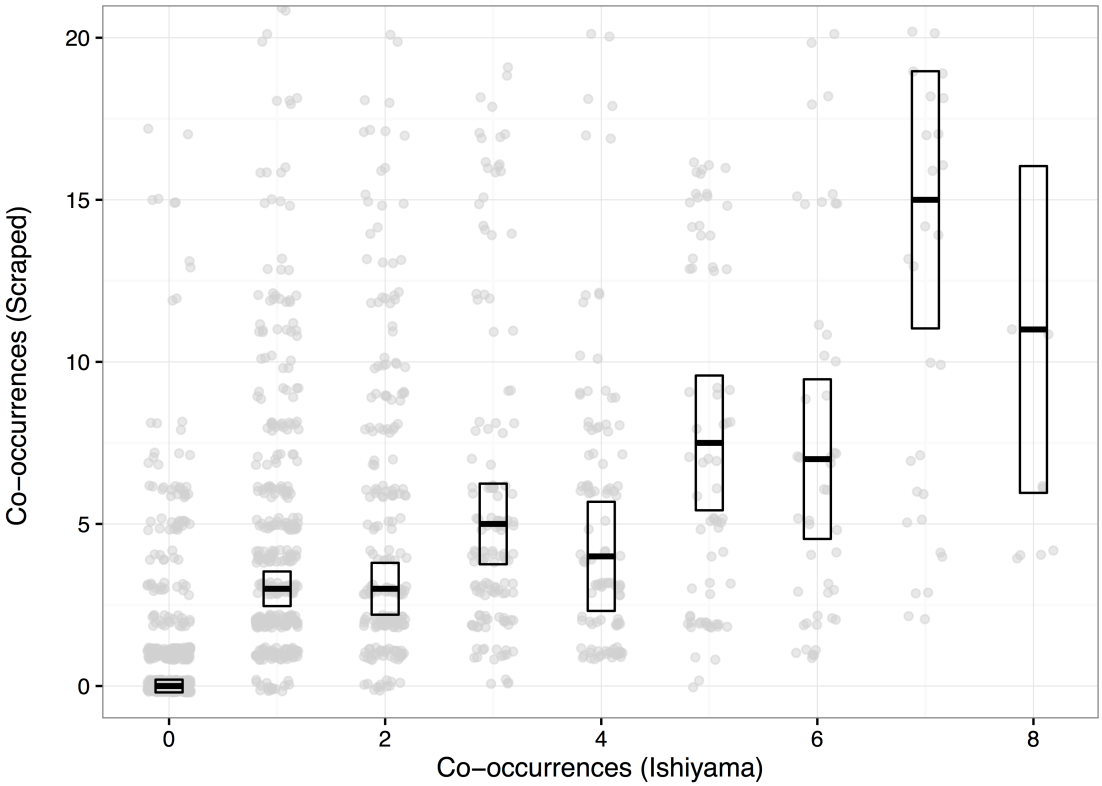


*Figure 2: Validating the proposed method with the Ishiyama (2014) North Korean elite network. Co-occurrences from the Ishiyama data (x-axis) are plotted with co-occurrences from the proposed method (y-axis). Medians with 95% confidence bands are included for each value of co-occurrences in the Ishiyama data (restricted to fewer than nine co-appearances with Kim Jong Un). Points are jittered to help alleviate overplotting.*

choosing date ranges or site restrictions—researchers should report results of different keyword combinations as a sensitivity analysis. For the US Senate Press Event network, for example, I am interested in the correlation between the scraped network and the existing network, so below I report the variation in these correlations across different network data based on different keyword searches.[4]

Figure 3 plots the correlation between the two networks as in the prior example. I apply a logarithmic transformation to the Google hits measure given it is skewed by a number of prominent senators in the sample (Barack Obama, Hillary Clinton, and others). The positive relationship between the two samples offers initial support for the method's conceptual validity, as senators occurring more frequently together at press events also appear more frequently together at social/political functions such as ceremonies and fundraisers. A log-log regression of the Google hits measure and press events measure indicates further support for the similarity of both measures: the coefficient of 0.84 (SE = 0.05, $t = 16.48$) suggests a nearly one-to-one relationship to co-appearances at social/political functions, though there is variation in the magnitude (but not statistical significance) of this relationship across different search keyword refinements (Table 1).

[4]Note that keyword variations may not always apply, as in the North Korea validation example above.

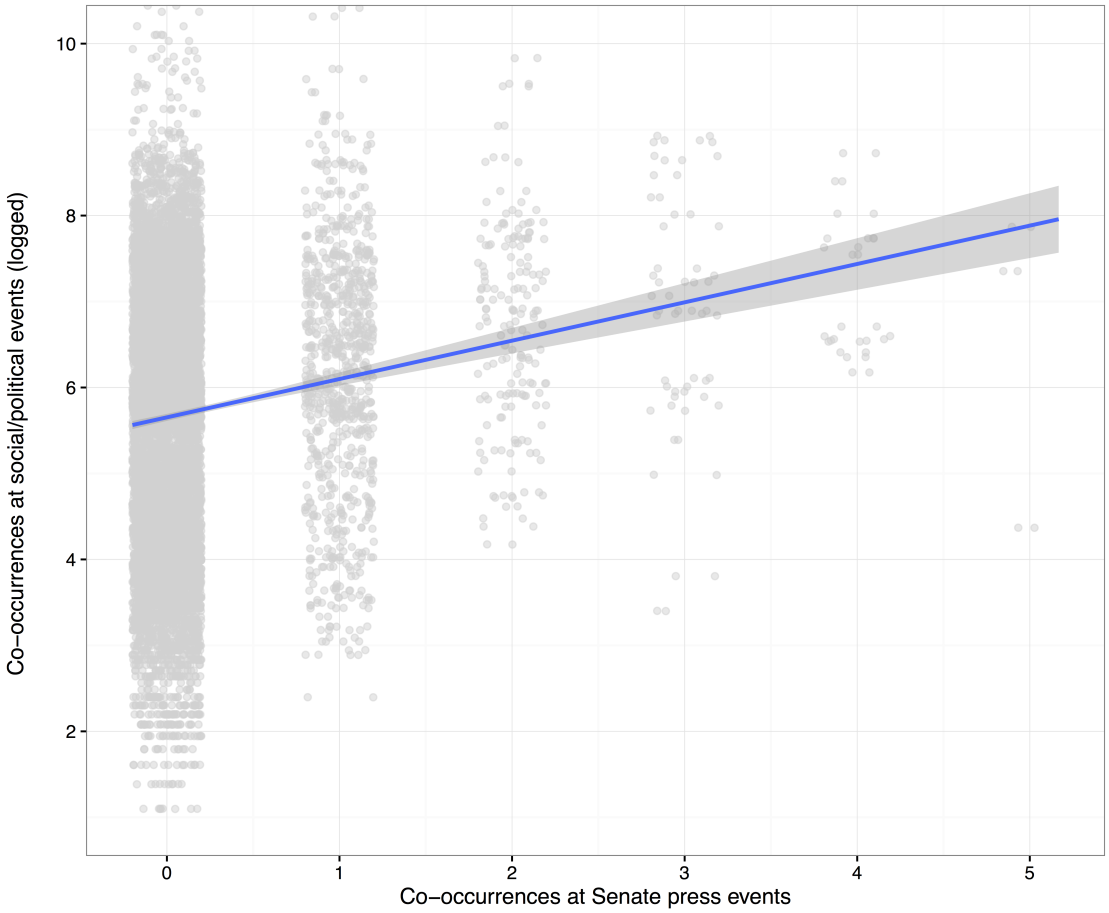


*Figure 3: Validating the proposed method with the Desmarais et al. (2015) senate press events network. A linear fit with a 95% confidence band is shown to compare co-occurrence in senate press events (x-axis, not logged) to co-occurrence in social/political events (y-axis, logged). The sample includes all possible combinations of 102 senators, with a total effective sample is 5,151 undirected dyads, represented in the plot by gray points. Co-occurrences at press events from Desmarais et al. (2015) range from 0 to 37, but for visual posterity I have omitted from the plot any points above 5 press event co-occurrences; this reduces the sample from 5,151 to 5,144. Including these 7 omitted points does not substantively change the result, given these pairs also have high public co-occurrences as measured via Google searches.*

Interestingly there is high variance in co-occurrences using Google hits for dyads with zero co-occurrences at press events, reflecting senators who appear at varying degrees of frequency at social events but never at senate press events. One such pair is Russ Feingold and Hillary Clinton who, despite not attending a single press event together, attended over 200 social events from 2007 to 2009. Indeed, Clinton rarely attended senate press events (a total of 4 during the 110th Congress) but was unsurprisingly active in the beltway social scene.

This variance between the two measures could be attributed to measurement error of the proposed technique as well as the "zero edges" problem discussed in the main text. But part of the difference between measures could also relate to construct differences: compared with

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Google hits, logged | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| Press events, logged | 0.841*** | 1.215*** | 0.156*** | 0.276*** | 0.219*** |
| | (0.051) | (0.067) | (0.018) | (0.025) | (0.020) |
| | t = 16.483 | t = 18.084 | t = 8.879 | t = 11.236 | t = 10.872 |
| Observations (undirected dyads) | 10,302 | 10,302 | 10,302 | 10,302 | 10,302 |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

*Table 1: OLS correlation between frequency of appearing at Senate press events measured by Desmarais et al. (2015) and by scraping co-occurrences. Models differ in keyword restrictions for the scraping algorithm:*

1. `gala+OR+ceremony+OR+fundraiser+OR+groundbreaking+event`
2. `gala+OR+ceremony+OR+press+event`
3. `press+event`
4. `gala+OR+dinner+OR+fundraiser`
5. `gala+OR+ceremony+OR+fundraiser+OR+groundbreaking+OR+event`

public co-occurrences, ties at Senate press events likely reflect ideological similarities more than ties fostered at social events, as the Clinton example illustrates (that is, if Feingold is less socially active for whatever reason). Nonetheless, as with the North Korea example above, researchers using the proposed technique should be cognizant of the sources of potential measurement error, as well as considerate of whether public co-occurrences at political and social events accurately capture the concept being studied.

Without the full sociomatrix of existing network data, one way to assess the accuracy of data from the proposed method is to calculate correlations of individual-level network measures of centrality. Centrality measures indicate how "important" given actors are within the network. The more ties an individual has to others in the network, the more likely that he/she is "central" within the population. The simplest measure of centrality is referred to as "degree centrality": in an undirected network, for each actor $i$ degree centrality is calculated as the total number of connections with all other actors in the network (formally this is given by $\sum_{j \neq i} x_{ij}$). A more robust measure is "Eigenvector centrality" which effectively captures not just how many individuals an actor is tied to but how important each of those individuals is within the network. Indeed, this is similar to the algorithm by which Google assigns "pagerank" for how search results are ranked and sorted. In addition, it is common for published work on network analysis to report centrality tables within the main text, while it is less common that the full sociomatrix is available to the public.

For three existing datasets without publicly available sociomatrices, I construct network data using the proposed algorithm and compute centrality scores for each node in the network. Figure 4 plots Eigenvector centrality scores from these data versus centrality scores from existing data. I find moderate-to-high correlations between network measures of centrality for each of the three samples. The lowest correlations come from comparisons to

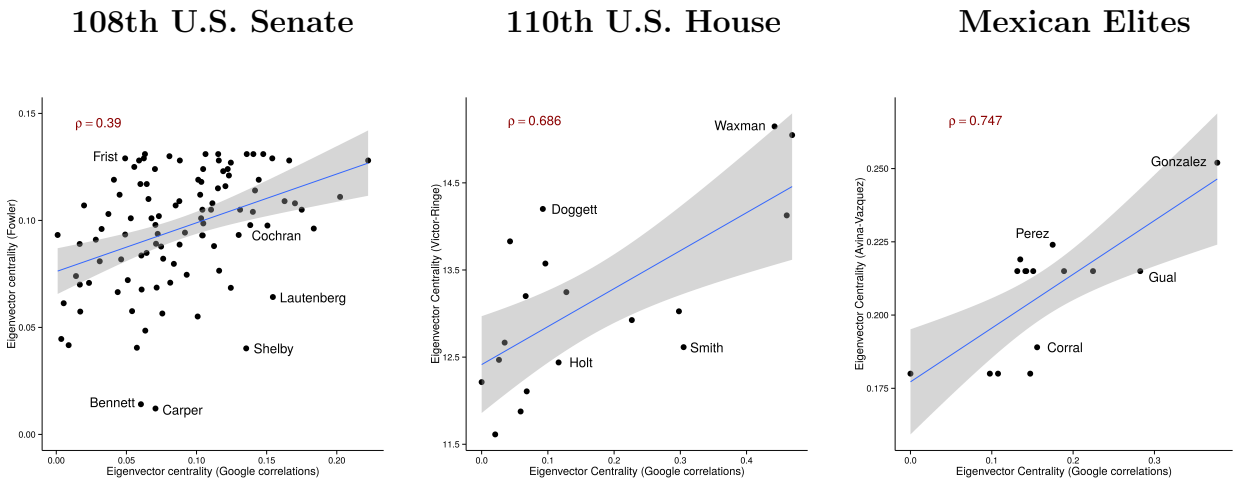


*Figure 4: Correlations of network centrality between data collected using scraped public co-occurrences versus data collected in previous research. A linear fit with a 95% confidence band is shown for each sample, along with the Pearson correlation printed in the top left of each plot.*

senate co-sponsorship data, where the correlation for Eigenvector centrality is 0.39. This is expected given the stark differences in how social ties are measured: co-sponsoring a bill is not likely to reflect the strength of social ties between senators in public political events.[5]

Senator Bill Frist is one such example of a senator who actively co-sponsors but seems to appear at few political events with other senators. Frist is ranked as highly central using co-sponsorship measures but in terms of co-occurrence at political events, he is relatively peripheral. The opposite pattern is exemplified by Senator Richard Shelby, who appeared frequently at political events co-attended by other senators, resulting in high centrality in terms of scraped public co-occurrences. Yet he co-sponsored relatively few bills (43, sample average = 78) and those he co-sponsored were typically with the same five Republican senators in the Deep south, thus resulting in low centrality by Fowler's measure.

Centrality correlations are slightly higher when comparing data from the proposed method to the Victor and Ringe (2009) data on representatives from the 110th Congress. Here I have restricted the House dataset given lack of data on the full 435 individuals in the sample. Instead, I analyze the top 20 most central representatives based on Table 3 in Victor and Ringe (2009, 756), and create a network for these 20 individuals' possible connections to all 435 individuals (the total size is thus $8,700$ possible dyads).[6] The resulting centrality correlations between the Victor-Ringe data and the network data created using the proposed algorithm are moderately high at 0.69 but note that much of the correlation is driven by high-profile representatives such as Henry Waxman.

[5]As Kirkland and Gross (2014) have noted, co-sponsorship is a relatively costless activity and senators may co-sponsor bills written by senators they may not necessarily be closely tied to socially. In addition, Victor and Ringe (2009) argue that co-sponsorship networks are conceptually distinct from social networks based on their analysis of caucus and organizational networks within Congress.

[6]Note that with the full network – if it were available – the resulting number would be 94,000 pairwise combinations. This could be scraped using the Google API but would require a longer time-frame (10 days) to do so, given that the Custom Search API is limited to 10,000 searches per day.

The highest correlations come from the Mexican board of directors data at 0.75. It is somewhat to be expected that of the three datasets without publicly available sociomatrices, there would be the most congruence in data collection techniques when looking at a network of high-profile business leaders. Given the high premium placed on networking in the executive community, board members of the same corporations are expected to interact with one another in social events and other professional situations (Carpenter and Westphal, 2001). In this sense, there is reasonable overlap between a measure of social ties based on co-occurrence at events and a measure of social ties based on co-membership on executive boards.

Taken together, these comparisons suggest that the method is best suited to collecting network data on elites where ties are conceptually based on social interactions as opposed to behavioral similarities. Valididations of the proposed scraping technique with the Fowler (2006) and Victor and Ringe (2009) data on congressional behavior (co-sponsorships and memberships in the same caucuses) are notably weak. The comparison to the Mexican board membership network provides some support for the validity of the proposed method, although it is difficult to make strong conclusions of conceptual accuracy with only 23 data points (given the lack of a full dyadic sociomatrix). On the other hand, validations with the Ishiyama (2014) and Desmarais et al. (2015) networks offer suggestive evidence for the conceptual accuracy of the scraping approach. The North Korean network in particular highlights the validity of the approach, with some measurement error notwithstanding, perhaps reflecting the appropriateness of the technique to elite public co-occurrence networks in authoritarian countries when compared to democracies such as the US and Mexico.

## Sampling random pages to ensure accuracy of search terms

A key step in the algorithm to ensure accuracy is to parse randomly sampled pages to determine if the resulting pages indeed capture co-occurrence at events. This is done by randomly sampling the JSON through-put that serves as an intermediary of the scraping procedure performed in the `perl` code shown below. Specifically, each search using the Google Search API runs through a JSON file with page names, URLs, and two-line snippets from the top 500 results. A section of one such file is shown in Figure 5. By looking at the two-line snippets in particular, the researcher can determine whether or not the page is appropriate to the search. *Note that the two-line snippets are for diagnostic use only – the actual search parses through the entire text in each article (and can process more than 500 results).*

This example shows the results from a search for "Harry Reid" and "Barbara Boxer" (two Democratic senators in the 109th U.S. Congress) with keyword restrictions of `groundbreaking` and `site:politico.com`. The result shown here (at the bottom of the file) indicates a page from *Politico* about Boxer's "groundbreaking idea" on a committee with Harry Reid which may have nothing to do with the co-occurrence of both individuals at a groundbreaking event. The solution in this context is to specify more relevant keywords using Boolean terms and quotations, such as (`'groundbreaking event'+OR+'groundbreaking ceremony'+OR+'hospital groundbreaking'+-'groundbreaking idea'`).[7] In general, refining the keywords in an iterative manner is necessary to ensure accuracy of the algorithm.

Beyond keyword-related measurement issues, there is also the concern of the well-known "page-counting problem" when using internet search engines such as Google (Lee et al. (2010); Clifton (2012)). When performing a Google search, users will see the "About _____ results" which is a rough approximation of the number of total pages containing or referring to the search keyword(s). The approximation, however, is typically *very* inaccurate, often over-counting the true number of results by a factor of 1,000. For example, a simple search for "Harry Reid" and "Barbara Boxer" — without any site restrictions or additional keywords — yields "About 320,000 results" but after clicking through to the last page of search results, this number dwindles to "About 349 results".

I overcome this page-counting problem by eschewing the collection of page hits via web-scraping of the html results page. Instead, I capture the total number of page hits based on the JSON through-put (as shown in Figure 5) provided by the Google Custom Search API. This approach provides an accurate count of total pages provided the total number of such pages is below 500, which is true for all pair-wise searches (with site restrictions and keywords) conducted in this paper. For searches that will likely result in more than 500 page results, the researcher can turn to scraping algorithms that will "click-through" to the final page of results and record the total number of resulting page hits. An additional option is to use sampling methods to generate "sample hits" based on a random sample of web pages from the first three to four pages of results — these initial pages theoretically represent the "best" results from a given search following the prioritization of pages established by the Google PageRank algorithm (Brin and Page (1998)).

[7] In Google searches, the 'NOT' Boolean term is operationalized with a '-' symbol.

```
"queries": {
  "request": [
   {
    "title": "Google Custom Search-Harry Reid Barbara Boxer groundbreaking site:politico.com",
    "totalResults": "9",
    "searchTerms": "Harry Reid Barbara Boxer groundbreaking site:politico.com",
    "count": 9,
    "startIndex": 1,
    "inputEncoding": "utf8",
    "outputEncoding": "utf8",
    "safe": "off",
    "cx": "012212847246781745017:u7disproqb8"
   }
  ]
 },
 "context": {
  "title": "senate"
 },
 "searchInformation": {
  "searchTime": 0.125084,
  "formattedSearchTime": "0.13",
  "totalResults": "9",
  "formattedTotalResults": "9"
 },
 "items": [
  {
   "kind": "customsearch#result",
   "title": "Boxer Tackles Challenge of Preserving Earth for Future Generations ...",
   "link": "http://www.politico.com/news/stories/0307/3184.html",
   "displayLink": "www.politico.com",
   "snippet": "Mar 19, 2007 ... Barbara Boxer (D-Calif.) ... it was really a very groundbreaking idea I had ... That's why I opened the microphone up to all my colleagues, and every other week, the chairmen meet at the call of (Senate Majority) Leader Harry Reid (D-Nev.), and we keep each other informed on the progress that's being made...",
   "cacheId": "uWMmAmpX728J",
   "formattedUrl": "www.politico.com/news/stories/0307/3184.html",
   "htmlFormattedUrl": "www.politico.com/news/stories/0307/3184.html"
}
```


*Figure 5: Example JSON through-put file using the Google search algorithm for a site search of U.S. Senators Barbara Boxer and Harry Reid with the keyword "groundbreaking" and a domain restriction to politico.com.*

## Modeling the Nigerian oil elite network using ERGM

The ERGM approach offers a model that estimates the effects of node-, dyad- and network-level covariates on social ties, given as:

$$P(Y = y|X = x, \eta) = \frac{\exp\left[\eta * g(y, x)\right]}{c(\eta, \mathcal{N})}, \qquad y \in \mathcal{N}$$

where $P(Y = y)$ is the probability of observing the network given by the data $(X, \eta)$; $\mathcal{N}$ is a set of possible networks; $\eta$ is a vector of parameters of interest; $g$ is a vector valued function; and $c(\eta, \mathcal{N})$ is a normalizing constant to ensure a finite integral.

Using this specification, I test whether board appointments to NNPC and social connectivity are correlated even when controlling for other individual- and network-level attributes. The former include individual popularity (self hits), being from the same region/province (regional homophily), or sharing the same ethnicity as the president (for the 2015 sample, I do not have data on geographic origins of each individual in the network). The latter includes network density (edges). Here, the dependent variable is connectivity — the presence and strength of a tie between two given individuals in the network. I construct two dependent variables based on co-occurrence as a measure of social ties. The first is binary: 1 if two individuals have attended the same political events together and 0 otherwise. The second variable is a discrete count of co-occurrent events between two individuals.

I test the hypothesis of connectivity and board appointments by including a dummy variable for whether an individual was ultimately appointed to the NNPC board of directors. If the coefficient on this term is statistically significantly greater (less) than zero, then I infer a positive (negative) correlation between political connectivity and board appointments.

ERGMs where nodal covariates serve as dependent variables are not well defined and may provide unstable coefficient estimates (Fellows and Handcock, 2013). Though this is a somewhat "roundabout" way of testing the determinants of board appointments — since appointments here are an independent variable — the approach can nonetheless estimate a *correlation* between connectivity and appointments, while importantly still accounting for the relational nature of the data. One alternative approach is to model the board appointments using a classical logistic regression model and use measures of network centrality as covariates: this would more accurately reflect the direction of the theoretical relationship that connectivity influences appointments. It should be noted that the non-network logit specification is the most commonly used approach in the literature, despite that it does not account for the endogeneity of the network and nodal characteristics.

More complex approaches involve the use of, for example, stochastic blockmodels or bipartite network models, or the largely untested Exponential family Random Network Model (ERNM) which does reflect the endogeneity of nodal characteristics and network values (Nowicki and Snijders, 2001; Karrer and Newman, 2011; Koskinen and Edling, 2012; Fellows, 2012; Fellows and Handcock, 2013). The former group of models have strong assumptions about the exogeneity of the outcome measure—in this case political appointments—to the network, an assumption which could in theory be relaxed with significant modifications to the model specification. Since my goal here is simply to estimate the correlation between connectivity and board appointments, it is outside the scope of the study to apply these

| | *2012* | | | *2015* | | |
|---|---|---|---|---|---|---|
| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
| Board appointee | −0.41 | 0.31 | 0.00 | 0.63** | −0.27 | −0.51*** |
| | (0.33) | (0.37) | (0.29) | (0.22) | (0.33) | (0.06) |
| Network density | −3.00*** | −5.23*** | −4.89*** | −2.66*** | −8.99*** | −12.49*** |
| | (0.22) | (0.43) | (0.33) | (0.19) | (0.61) | (0.10) |
| Cabinet homophily | 0.58* | 1.35*** | 0.84** | 1.43*** | 0.54 | −0.13** |
| | (0.28) | (0.36) | (0.28) | (0.21) | (0.31) | (0.05) |
| Region homophily | | −0.08 | −0.17 | | | |
| | | (0.38) | (0.31) | | | |
| Pres co-ethnic | | 0.65* | 0.75*** | | | |
| | | (0.26) | (0.22) | | | |
| Google self-hits | | 0.39*** | 0.38*** | | 0.82*** | 1.15*** |
| | | (0.04) | (0.03) | | (0.06) | (0.01) |
| AIC | 553.91 | 447.96 | | 1061.51 | 466.47 | |
| BIC | 569.48 | 479.10 | | 1077.08 | 487.23 | |
| $N$ (dyads) | 1326 | 1326 | 1326 | 1326 | 1326 | 1326 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

*Table 2: Exponential Random-Graph Models of social connectivity among Nigeria's oil elite. The table shows results from three specifications: cols. 1–2, 4–5: ERGM; cols. 3, 6: ERGM-count.*

classes of models to the data; instead I focus only on the "roundabout" ERGM approach and the parsimonious logit model.

Results from ERGM model specifications are presented in Table 2. In models 1–2 and 4–5, the dependent variable is binary (either a tie exists between two individuals or not) and the specification is the standard ERGM. In models 3 and 6, I apply the ERGM-count specification with a discrete count measure of ties. Model diagnostics for the specification including transitivity (column 2) indicate a reasonable fit of the ERGM to the data, with the exception of nodes with 5 to 6 edge-wise shared partners.

Across all model specifications, there very little evidence of a positive correlation between board appointments and social ties. For example, the coefficient for board appointee in model 3 (0.00) implies that there is a 50% probability of a tie forming between two individuals if one is a future board appointee, which is precisely the baseline probability if ties formed between these individuals by random chance. Once self-hits are controlled for in the 2015 sample, the coefficient for board appointment is negative — suggesting that ties between individuals where one is a future appointee are less likely than ties between non-appointees.

One interesting pattern revealed by the network analysis is that there is evidence of clustering in the 2012 sample. Specifically, the coefficient on the cabinet homophily term is positive and large in substantive terms, suggesting that if two individuals are both cabinet ministers, they are very likely to be connected (between 64% and 79% probability of tie

| | *2012* | | *2015* | |
|---|---|---|---|---|
| | Model 1 | Model 2 | Model 3 | Model 4 |
| Intercept | −0.96 | −0.89 | −1.81** | −1.68** |
| | (0.66) | (0.65) | (0.58) | (0.56) |
| EV centrality | 2.96 | | 1.52 | |
| | (7.87) | | (9.43) | |
| Degree centrality | | 0.11 | | 0.02 |
| | | (0.14) | | (0.02) |
| Pres co-ethnic | 1.22 | 1.25 | | |
| | (1.11) | (1.10) | | |
| South dummy (region) | −0.31 | −0.43 | | |
| | (0.99) | (1.01) | | |
| Google self-hits | −1.29 | −1.69 | −0.06 | −0.25 |
| | (1.05) | (1.12) | (0.36) | (0.29) |
| AIC | 51.80 | 51.21 | 47.06 | 46.09 |
| BIC | 61.55 | 60.97 | 52.91 | 51.95 |
| Log Likelihood | −20.90 | −20.61 | −20.53 | −20.05 |
| Deviance | 41.80 | 41.21 | 41.06 | 40.09 |
| $N$ (nodes) | 52 | 52 | 52 | 52 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

*Table 3: Logit models of social connectivity among Nigeria's oil elite (standard errors in parentheses). The dependent variable is appointment to the NNPC board (0 = no, 1 = yes).*

formation based on estimates from models 1 and 2). Indeed, looking at the network graph in Figure 1 in the main text, we can see this clustering of NNPC officials in the nodes to the right of Goodluck Jonathan (in red). Controls for regional clustering, however, do not show evidence that there is grouping by region (province) of origin, though there is evidence that the President's co-ethnics are likely to be socially tied.

A brief note on the ERGM-count results: similar to a Poisson model, the dependent variable is a logged count of non-negative discrete values, so interpretation is easier after exponentiating coefficients. This gives statements of relative changes in the non-transformed counts of event co-occurrence among two given individuals. For instance, the −0.52 coefficient for board appointees in model 6 indicates that there is a predicted 40.5% decrease in the count of events co-attended by two individuals if one of them is a board appointee.

Conventional (non-network) logistic regression models show the same results as above. Table 3 provides estimates using logit models where board appointment is the dependent variable and two metrics of network centrality—Eigenvector centrality and degree centrality—are alternately used as independent variables, along with the same controls as before.[8] Again,

[8]Cabinet homophily is excluded due to singularities: in 2012 no cabinet members were appointed to the NNPC board.

there is little correlation between board appointments and co-occurrences as measured by network centrality. Nor is it the case that more popular individuals (those with higher self-hits) more likely to be appointed to the NNPC board.

## Perl Script for Google Custom Search API

The following code provides one framework for scraping Google using the Google Custom Search API, which requires used to register for an API key (https://cse.google.com/cse/). Searches above 100 queries per day require a licensed Google Custom Search API. Users should note that while scraping can be accomplished without the use of an API key and instead using proxy sites, this approach might violate Google's Terms of Service depending on frequency of usage and number of requests per use. As such, the use of proxies is not recommended. An alternative approach is to use the Google News Archive search though this restricts international newspaper searches to the post-2012 period.

The code below, which is currently being converted for use in an `R` package, can be modified to scrape co-occurrences using sites other than Google. This includes direct sources of information whose websites are built using Google custom search, such as Nigeria's *Vanguard*, or potentially using media databases such as LexisNexis.

```
use strict;
use warnings;
use LWP::UserAgent;
use URI::Escape;
use JSON;
my%configuration = (
# search API id goes here
  id => '',
# search API key goes here
  key => '',
# input two-column csv file with full list of n*(n-1)/2 pairs in network
    population
  input => '',
# output file name
  output => '',
# search criteria & restricted domain name(s)
  expression => '%s+and+%s+SEARCHCRITERIA+site:DOMAIN.COM',
# number of seconds to pause in between pairs
  sleep => 1
);

my$agent = LWP::UserAgent->new;
open( my$input, '<', $configuration{input} ) or die $configuration{input}.': '
    .$!;
open( my$output, '>', $configuration{output} ) or die $configuration{output}.'
    : '.$!;

# scraper set up
while( <$input> ){
# set up output edgelist based on pairs & scraped hits variable ('value')
  chomp;
  print $output $_;
  my$q = 0;
  my$f = 0;
  my@c = split( '' );
```

```
  my( $value, @value );
  for( my$i = 0; $i <= $#c; $i++ ){
    if( $c[$i] eq '"' ){
      if( $q == 0 ){
        if( exists $c[$i + 1] && $c[$i + 1] eq '"' ){
          $q = 1;
        }else{
          if( $f == 0 ){ $f = 1 }else{ $f = 0 };
        }
      }else{
        $value .= $c[$i];
        $q = 0;
      }
    }elsif( $c[$i] eq ',' && $f == 0 ){
      push @value, $value;
      undef $value;
    }else{
      $value .= $c[$i];
    }
  }
  push @value, $value if defined $value;
  print '1: '.$value[0].'   2: '.$value[1]."\n";

# scraping co-occurrences: hits for each pair & keywords ('expression')
  my$response = $agent->get(
# search engine API to use for scraping
    'https://www.googleapis.com/customsearch/v1?'.
# search API id
    'cx='.$configuration{id}.'&'.
# search API key
    'key='.$configuration{key}.'&'.
# setting URL based on pairs and keywords, avoiding non-URL friendly
   characters
    'q='.sprintf( $configuration{expression}, uri_escape( $value[0] ),
    uri_escape( $value[1] ) ).'&'.
# specifying what to collect; i.e., hits ('totalResults')
    'fields=searchInformation/totalResults'
  );
# setting up what to do with each result
  if( $response->is_success ){
# scrape from JSON intermediary ('response')
    my$data = decode_json( $response->decoded_content );
# collect hits from each JSON output
    my$hits = $$data{searchInformation}{totalResults};
# show progress of scraper, printing each unique URL and pair names
    print 'response: '.$hits.' result';
    $hits == 1 ? print "\n" : print "s\n";
    print $output ','.$hits."\n";
# show pairs with errors
  }else{
    print 'response: '.$response->status_line."\n";
    print $output ','."\n";
  }
# wait n seconds before moving to next pair
```

```
85    sleep $configuration{sleep};
  }
87  close $input;
  close $output;
89  # show completion of scraping
  print "done\n";
91  exit;
```

# References


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

Brin, S. and L. Page (1998). The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems 30*(1), 107–117.

Carpenter, M. A. and J. D. Westphal (2001). The strategic context of external network ties: Examining the impact of director appointments on board involvement in strategic decision making. *Academy of Management Journal 44*(4), 639–660.

Clifton, B. (2012). *Advanced Web Metrics with Google Analytics*. Wiley.

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.

Fellows, I. (2012). *Exponential Family Random Network Models*. Ph. D. thesis, University of California, Los Angeles.

Fellows, I. and M. S. Handcock (2013). Exponential-family Random Network Models. *arXiv preprint arXiv:1208.0121*.

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

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.

Karrer, B. and M. E. Newman (2011). Stochastic blockmodels and community structure in networks. *Physical Review E 83*(016107), 1–10.

Kirkland, J. A. and J. H. Gross (2014, January). Measurement and theory in legislative networks: The evolving topology of Congressional collaboration. *Social Networks 36*(1), 97–109.

Koskinen, J. and C. Edling (2012). Modelling the evolution of a bipartite network: Peer referral in interlocking directorates. *Social Networks 34*(3), 309 – 322. Dynamics of Social Networks (2).

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.

Nowicki, K. and T. A. B. Snijders (2001). Estimation and prediction for stochastic block-structures. *Journal of the American Statistical Association 96*(455), 1077–1087.

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.