**Title**
Google Correlations: New approaches to collecting data for statistical network analysis



**Author**
Mahdavi, Paasha

**Publication Date**
2014

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

# Google Correlations: New approaches to collecting data for statistical network analysis

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

**Paasha Mahdavi**

2014

Abstract of the Thesis

# Google Correlations: New approaches to collecting data for statistical network analysis

by

**Paasha Mahdavi**

Master of Science in Statistics

University of California, Los Angeles, 2014

Professor Mark Handcock, Chair

This thesis introduces a new method for data collection on political elite networks using non-obtrusive web-based techniques. One possible indicator of elite connectivity is the frequency with which individuals appear at the same political events. Using a Google search scraping algorithm (Lee 2010) to capture how often pairs of individuals appear in the same news articles reporting on these events, I construct network matrices for a given list of individuals that I identify as elites using a variety of criteria. To assess cross-validity and conceptual accuracy, I compare data from this method to previously collected data on the network connectedness of three separate populations. I then supply an application of the Google method to collect network data on the Nigerian oil elite in 2012. Conducting a network analysis, I show that appointments to the Nigerian National Petroleum Corporation board of directors are made on the basis of political connectivity and not necessarily on technical experience or merit. These findings lend support to hypotheses that leaders use patronage appointments to lucrative bureaucratic positions in order to satisfy political elites. Given that many political theories on elite behavior aim to understand individual- and group-level interactions, the potential applicability of network data using the proposed technique is very large, especially in situations where collecting network data intrusively is costly or prohibitive.

The thesis of Paasha Mahdavi is approved.

Qing Zhou

Ying Nian Wu

Mark Handcock, Committee Chair

University of California, Los Angeles

2014

*To my parents*

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgments

First thanks go to Mark Handcock, my committee chair, for his endless help with this project from its beginnings as an outline for STAT 218 to its current, finished form. His contributions were extremely helpful in all phases of the project, from improving the search algorithm to assessing cross-validity to modeling the correlates of the resulting network. I also acknowledge the help of my other committee members, Qing Zhou and Ying Nian Wu, for reading the thesis and offering their comments and suggestions. Thanks also to my faculty mentors in the Department of Political Science, in particular Michael Ross, Jeffrey Lewis, and James DeNardo, who greatly encouraged and helped me in my goals of completing a M.S. in Statistics while working on my Ph.D. in Political Science. Thanks to the department administrator, Glenda Jones, who has been instrumental in helping me navigate the Master's program these past two years. Without her help, I would not have been able to finish the thesis and the Master's.

Beyond my professional colleagues, this thesis could not have been completed without the help and support of my family and friends. Thanks especially to my brother, my sister, and my mother and father. Their emotional support throughout my graduate studies and my career has been constant, unwavering, and unconditional. I also could not have completed this project (and many others) without my once and future collaborator, Felipe Nunes, whose guidance throughout graduate school has been essential. Lastly, perhaps my greatest thanks and acknowledgements are to Megan – my best friend, my inspiration, and my star forever.

# CHAPTER 1

# Introduction

In the social sciences, researchers are currently equipped with a number of techniques for network data collection on individual actors, ranging from direct observation (ethnography), to surveys and interviews, to machine learning tools used with archival records (e.g., text analysis of Senate bills). These approaches are well-suited to the study of legislative politics (Fowler, 2006; Victor and Ringe, 2009; Kim, Barnett and Park, 2010; Alemán and Calvo, 2013; Kirkland and Gross, 2014), voter turnout in the U.S. (Bond et al., 2012; LaCour and Green, 2014), judicial politics (Fowler et al., 2007; Lupu and Voeten, 2012), party politics (Koger, Masket and Noel, 2010; Morgan, 2014), and interest group behavior (Box-Steffensmeier and Christenson, 2014). In particular, current approaches using archival records with machine learning techniques are well-suited to create network data based on legislative bills, court citations, and party or membership lists. Outside of these areas of study, however, it is considerably more difficult to apply these methods.

Consider the costs and feasibility of using existing methods to gather network data on, for instance, clientelism networks between politicians and voters in developing countries. Text analysis is difficult given there is no well-defined corpus of texts that can be used to infer ties (other than confidential vote-buying lists held by political leaders). Further, applying survey and interview methods may not only be extremely costly but also potentially dangerous to researchers. Enumerators who are surveying individuals about vote-buying in Iran, Myanmar or North

Korea are likely to be imprisoned, while interviewers of vote-buyers in war-zones such as Afghanistan, Central African Republic, and Syria are at risk of death or severe injury. Scholars engaged in the study of autocratic 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.

In this thesis, I provide a new method for gathering network data on political elites in costly and prohibitive contexts, such as in authoritarian countries or conflict-ridden societies. Specifically, I introduce a technique to assemble relational data based on a given individual list of political elites. For example, Avina-Vazquez and Uddin (2013) analyze the professional connectedness of Mexico's richest and most successful businessmen, as defined by their board membership in "Fortune 50" Mexican corporations. Given a list of individuals, 23 of whom are specified in the table on the left-hand side of Figure 1.1, the data collection goal is to assemble information on professional ties between any pair of the 23 individuals. The graph on the right-hand side of Figure 1.1 shows one possible network based on this list of individuals. Getting from a list of people to data on how they are tied to one another is the necessary first step in any standard network analysis on individual actors.

One way to measure social ties — that is, how individuals are connected to one another — is to capture the frequency with which individuals attend the same events, organizations, groups, or other social activities. This measure of "co-occurrence" is one possible method (in the next section I describe other methods) to observe what is otherwise a latent, or unobservable, characteristic — the presence of a social connection between individuals. Using this measure of social ties, I construct a web-based algorithm using the Google search engine to capture co-occurrence among a list of pre-specified actors at political events that are covered

by reputable online media sources. To glean social ties among U.S. senators, for example, I use the algorithm to observe how frequently senators attended the same political functions in Washington, D.C., together as reported by sites such as *Politico*, *The Hill*, *Wonkette*, and *The Washington Post*, among others. As I describe in the next section, network data on U.S. senators can be collected using a variety of existing tools that are both and feasible and cost-effective. In trying to collect this kind of information on oil elites in Nigeria, for example, existing tools are less appropriate and necessitate a new approach to network data collection.

This paper is structured as follows. In the next section, I describe existing techniques for network data collection in political science and I discuss the benefits and drawbacks of each approach. In the third section, I explain the proposed algorithm and I outline the process as it works in practice. In the fourth section, I test the conceptual accuracy of the algorithm by cross-validating with three existing datasets on political elites. In the fifth section, I apply the algorithm to gather data on the Nigerian oil elite and I provide a brief modeling exercise using the newly-created network data. In the sixth and final section, I conclude with a discussion of future improvements to the technique.

| | Company |
|---|---|
| Alberto Bailleres Gonzalez | Dine, S.A. |
| Alejandro Bailleres Gual | Grupo Nacional Provincial |
| Alejandro Paredes Huerta | Tecnica Administrativa Bal |
| Arturo Fernandez Perez | Bimbo |
| Carlos Orozco Ibarra | Grupo Alfa |
| Claudio Salomon Davidson | Grupo Palacio de Hierro |
| Claudio X Gonzalez Laporte | Grupo Alfa |
| Dolores Martin Cartmel | Tecnica Administrativa Bal |
| Eduardo Silva Pylypciow | Grupo Profuturo |
| Emilio Carrillo Gamboa | GMexico |
| Enrique Castillo Sanchez Mejorada | Grupo Alfa |
| Fernando Ruiz Sahagun | GCC |
| Fernando Senderos Mestre | Dine, S.A. |
| Joaquin Vargas Guajardo | Fundacion CMR |
| Jose Antonio Fernandez Carbajal | Bimbo |
| Jose Luis Simon Granados | Grupo Nacional Provincial |
| Jose Octavio Figueroa Garcia | Tecnica Administrativa Bal |
| Juan Bordes Aznar | Grupo Nacional Provincial |
| Rafael Alfonso Mac Gregor Anciola | Fresnillo Plc. |
| Raul Bailleres Gual | Industrias Penoles |
| Raul Obregon Del Corral | Bimbo |
| Tomas Lozano Molina | Corporacion GEO |
| Valentin Diez Morado | Grupo Alfa |

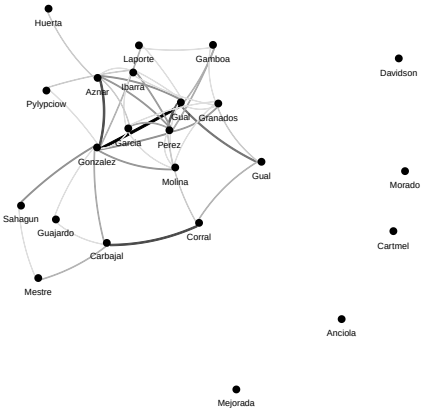


Figure 1.1: Network data on the Mexican business elite: Social ties among board members in "Fortune 50" companies in Mexico. The table on the left lists 23 individuals on prominent boards in Mexico-based corporations (Source: Avina–Vazquez and Uddin (2013)). The network graph on the right maps their connections using the proposed Google correlations technique of network data collection.

# CHAPTER 2

# Techniques for Network Data Collection

Political scientists have applied a variety of approaches to collect data for statistical network analysis on political actors, including questionnaires, observations, experiments, and archival record analysis. In this section, I review current techniques for network data collection and discuss some limitations when applying these approaches to collecting network data in authoritarian and war-torn contexts.

## 2.1 Current Approaches

Initial quantitative research of network analysis in political science relied primarily on archival records — political documents and texts — to collect relational data. Early network analysis of legislative influence quantified common interests among U.S. congressional representatives to create measures of social ties. Using legislative roll call texts, Fowler (2006) measures relational ties as the frequency of bill co-sponsorships across senators and congresspeople. Specifically, a directional tie exists between two individuals if one has co-sponsored a bill proposed by the other (see Alemán and Calvo (2013) for an application of this method in Latin America). Victor and Ringe (2009) instead measure ties as the frequency of co-membership in caucus organizations, using documents such as caucus lists and annual congressional reports of caucus memberships. Here, two senators are tied if they both are members of the same caucus group. A more dynamic source of information to create senate networks is proposed by Kim, Barnett and Park

(2010), who use senators' congressional websites to create a network based on the number of inlinks and outlinks on each members' site. A senator $i$ is directionally tied to another senator $j$ if his/her website contains an outlink to senator $j$.

Scholars working on party organizations — consisting not only of elected representatives but also candidates, interest group members, and others in the policy process — similarly use texts and official lists to create relational data. In a multiparty legislative context, Morgan (2014) creates dynamic network data on Polish legislators using publicly-available party lists as they evolve over time. Beyond party membership, Koger, Masket and Noel (2010) use proprietary lists of the transfers of donor information between political organizations to create relational data on "extended party networks" (see also Cohen et al. (2009)). For example, an individual donating money to both 2004 presidential candidate Howard Dean and House speaker Nancy Pelosi would indicate a tie between Dean and Pelosi. Within the context of judiciary networks, Lupu and Voeten (2012) use court case citations to construct relational data on judges in the European Court of Human Rights, where two judges citing the same court case are considered to be tied to one another.

The Mexican board of directors data presented above in Figure 1.1 is similarly constructed using membership lists. Here Avina-Vazquez and Uddin (2013) use records of board membership in "Fortune 50" publicly-traded corporations to construct network data for board members across the country. This technique follows from earlier sociological work on CEO networks in the United States by Galaskiewicz (1985).

For network analysis of non-elected elites such as voters and local leaders, researchers have often turned to using more traditional methods of gathering network data common in the disciplines of sociology and anthropology. By conducting surveys and interviews, one can infer social ties via self-identification of individuals to which a subject is connected. The classic study of American political

elites by Moore (1979), for instance, uses the survey-based *American Leadership Study* along with interviews of leaders of political organizations to characterize the degree of integration within and across local political elites. More recently, Grossman (2014) collects network data on rural Ugandan community elites by distributing questionnaires and conducting interviews with community leaders.

Moving beyond self-reported data, one can employ techniques based on respondent-driven sampling (Gile and Handcock, 2010; Lange, 2014). Beginning with an initial sample of respondents, the researcher solicits additional respondents by asking initial subjects to nominate or recruit others to participate in the study or questionnaire. While common in sociology and public health network studies, few political scientists use this approach. The exception is a study by LaCour and Green (2014) who map voter networks in east Los Angeles county using an initial sample augmented with a "snowball" sample of respondents recruited by subjects in the initial sample. Here, a voter in the baseline sample is asked at the end of an online survey to nominate five of his/her friends to take the survey as well, with monetary incentives offered to the subject for his/her friends who successfully complete the survey.

Another technique for subject-driven network data is the use of online interactions to measure social ties in a given population. Work by Bond et al. (2012) and Jones et al. (2013) uses proprietary Facebook data to infer social ties among a sample of the voting-age population (VAP) in the United States. Both studies use Facebook interactions — messaging, wall-posting, sharing, and liking — to measure social tie strength between individuals in the sample. The more two individuals interact online, the more they are inferred to be socially tied in the non-online world.[1]

Arguably more obtrusive a technique is direct observation through ethno-

[1] Jones et al. (2013) cross-validate the online interactions data with survey responses from a reduced sample of users.

graphic study. With this approach, one can identify social ties between individuals through direct observation of the behavior and habits of a given population (Wasserman and Faust, 1994). The famous "Sampson's Monks" network dataset is described as being collected through years of direct observation of the behavior of a group of monks (Sampson, 1969). In analyzing the formation of nomadic clan networks, White and Johansen (2005) apply this technique to assemble a relational dataset by directly observing social interactions between nomadic leaders in rural Turkey. An earlier study by Singerman (1995) (that does not explicitly use network analysis) similarly applies the ethnographic approach to gather information on how poor residents in Cairo use informal political networks and government subsidies.

## 2.2 Limitations

While the techniques for network data collection in political science have yielded fruitful relational datasets thus far, there are considerable limitations in using these approaches in attempting to create network data in the developing world, particularly in non-democracies and war-torn countries. Three limitations in particular must be addressed.

The first is the cost of using existing approaches. Obtrusive data collection techniques such as surveys, RDS, and direct observation can be financially costly to implement for tough-to-reach populations. Interviews and ethnographic observation can be additionally costly in terms of time spent collecting data. Importantly, for all obtrusive techniques, 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.

Second, obtrusive techniques can also suffer from respondent-induced measurement error. Though measurement error occurs in nearly all social network research

(Wasserman and Faust, 1994), data collection designs that rely on self-reports are particularly vulnerable to discrepancies between observed and true ties (Holland and Leinhardt, 1973). Specifically, the validity of self-reported social ties are often called into question based on potential respondent bias in whom subjects choose to report as their friends or in the strength of reported social ties.[2]

Current machine learning methods based on archival records or texts solve these problems by capturing behavioral rather than reported ties. Yet these techniques require an identifiable and consistent corpus of text to analyze (e.g. legislative bills, court citations, party lists) that limits these techniques to studies of presidents, legislators, party affiliates, and judges. Thus a third limitation of existing approaches to network data collection is one of scope. To collect data on oil elites in Nigeria, as in the application below, there are no such lists or archival records. The same is true for trying to collect network data on members of the military regime in Egypt, the extended monarchy network in Saudi Arabia, or the clientelism network in Mexico under the PRI regime.

[2]It should be noted that RDS and snowball sampling designs can help to reduce self-reported measurement error given the additional burden of recruitment or nomination of friends as opposed to simply listing or reporting names of friends.

# CHAPTER 3

# Proposed method: Google Correlations

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

This type of network 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). The use of affiliation networks in sociology dates back to 1941 with the study by Davis, Gardner and Gardner (1941) on gathering relational data on the social activities of eighteen Southern women through the use of newspapers and interviews to record how often the subjects attended the same social events. This approach is less common in political science studies of individual actors,[1] though legislative co-sponsorship, caucus membership, and the Mexican board of directors networks can be considered as affiliation networks. But 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.

[1]Note, however, that nearly all network analysis studies in the sub-field of international relations can be considered affiliation networks, given that co-occurrence of countries within given organizations, alliances, or trade agreements is the primary source for measuring ties between countries. For a review of these studies, see Hafner-Burton, Kahler and Montgomery (2009).

Lee et al. (2010) introduce such a technique using web-based search engines to collect this type of data in practice. The use of web-based search engines relies on the logic that the more often two individuals co-appear in the same news articles, blogs, webpages, etc., the more likely it is that these two individuals are closely related when compared to two random counterparts in a given sample. Lee et al. (2010) term this co-occurrence as the "Google correlation" between two individuals in a network. Specifically, it is defined as "the number of pages searched using Google when the pair of members' names ... is entered as the search query" (Lee et al., 2010, 1).[2] Note that, importantly, the frequency of appearing in news articles together is conceptually correlated with but not strictly equal to the frequency of interactions.

I build on the approach introduced by Lee et al. (2010) to apply the technique to collect data on political 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 machine learning 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 in online media reports. For example, to capture co-occurrence at political fundraisers in Washington, D.C., I use keywords such as "fundraiser", "campaign event", 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`.

Using this approach, the sociomatrix — a $n \times n$ matrix where each cell contains the value of a social tie between actors $i$ (rows) and $j$ (columns) — is constructed 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

[2]The authors also use "additional word(s)" to refine searches to specific time periods, name qualifiers (e.g. "professor", "senator") and other restrictions.

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. In this way, 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. The script is written in `perl` with code available in the appendix.[3] 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 (Grimmer and Stewart, 2013). Here, the researcher combs through randomly sampled pages to

[3]Searches above 100 queries per day require a licensed Google Custom Search API. An alternative approach is to use the Google News Archive search, following Chadefaux (2014), though this restricts international newspaper searches to the post-2012 period. In the appendix, I also provide `python` code which uses proxies for iterative searches.

determine if the resulting pages 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 in the appendix. 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 3.1. By looking at the two-line snippets in particular, the researcher can determine whether or not the page is appropriate to the search.

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'`).[4] 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 (Berthon, Pitt and Prendergast (1997); Lee et al. (2010); Clifton (2012)).[5] 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.

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

[5] See also "Google Inconsistencies" (2003) http://www.searchengineshowdown.com/features/google/inconsistent.shtml. Accessed on 16 Sep 2014.

```
"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 3.1: 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.

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 3.1) 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

must turn to *ad hoc* solutions such as "capping" search results above 1,000 (Lee et al. (2010)) or designing 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)).

# CHAPTER 4

# Conceptual Accuracy: Assessing Validity of the Proposed Method

Like any method of network data collection, the proposed approach using Google correlations is subject to measurement error. Three potential sources of error abound. 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, 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. Third, 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).

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. To assess the method's conceptual accuracy, I validate Google correlations results using three existing sets of network data:

- Co-sponsorships in the 108th U.S. Senate (Fowler, 2006);

- Caucus memberships in the 110th U.S. House (Victor and Ringe, 2009);

- Elite Board Members in Mexico (Avina-Vazquez and Uddin, 2013).

If measurement error using the proposed algorithm is high (for reasons specified above), then the network data resulting from Google correlations should not conform with existing network data. One way to assess the accuracy of the data is to calculate correlations of individual-level network measures of centrality for both sets of data. 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.

For each of the three datasets above, I construct network data using the proposed Google correlations algorithm and compute centrality scores for each node in the network. In Figure 4.1, I plot centrality scores from these data versus centrality scores from existing data. In the top half of the plot, I compare Eigenvector centrality scores, whereas in the bottom half I compare degree centrality scores.

Overall, I find moderate-to-high correlations between network measures of centrality based on using my algorithm versus using existing network data. The lowest correlations come from comparisons to Fowler's co-sponsorship data, where the correlations for degree and Eigenvector centrality are 0.39 and 0.51, respectively. This is expected given the stark differences in how social ties are measured. Whereas Fowler (2006) considers senators to be tied if they have co-sponsored the same bill (or if one is the author of the bill and the other is a co-sponsor), I consider actors tied if they have attended the same political events together.

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.[1] 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.

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, identified in the top left of the Eigenvector centrality comparison, 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, identified in the bottom right of the same graph. Shelby appeared frequently at political events co-attended by other senators, resulting in high centrality in terms of Google correlations. 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.

Before turning to comparison plots for data from Victor and Ringe (2009) and Avina-Vazquez and Uddin (2013), it is interesting to note that there is an apparent "ceiling effect" for Eigenvector centrality using the Fowler (2006) data. Specifically, no individual has a centrality score higher than 0.131, while there is no such ceiling for Eigenvector centrality scores using the Google correlations data. This may be an artifact of the assumptions in Fowler's centrality calculations, though it is unclear why there is no such ceiling when looking at degree centrality (where one might expect a ceiling effect given the finite number of ties an individual can form in the network).

Centrality correlations are higher when comparing data from the Google cor-

[1] Kirkland and Gross (2014) also find reason to question the use of co-sponsorship as a social tie measure given evidence that it is highly sensitive to measurement choices such as time, strength of tie, and direction of ties based on authorship vs. co-sponsorship.

relations method to the Victor and Ringe (2009) data on representatives from the 110th Congress. Here I have restricted the House dataset given the intractability of searching over $\binom{435}{2} = 94,395$ individuals to create a full network. Instead, I create a network for the top 20 most central representatives and their possible connections to all 435 individuals (the total size is thus $8,700$ possible dyads). The resulting centrality correlations between the Victor-Ringe data and the network data created using the proposed algorithm are moderately high at 0.69 and 0.60, respectively for degree and Eigenvector centrality.

The highest correlations come from the Mexican board of directors data. Here the degree and Eigenvector centrality measures across both datasets have correlations of 0.75 and 0.87, respectively. It is somewhat to be expected that of the three datasets, 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.

These three comparisons indicate that the proposed Google correlations algorithm is reasonably accurate in creating network data based on co-occurrence of individuals at social and political events. This is not to say that there is no measurement error in the approach. Refining the algorithm and cross-validating with existing data where available is important to improving the conceptual accuracy of the proposed method. The burden is thus on the researcher to assess the validity of the algorithm's output with other sources (qualitative or quantitative) based on the characteristics of the population of interest.

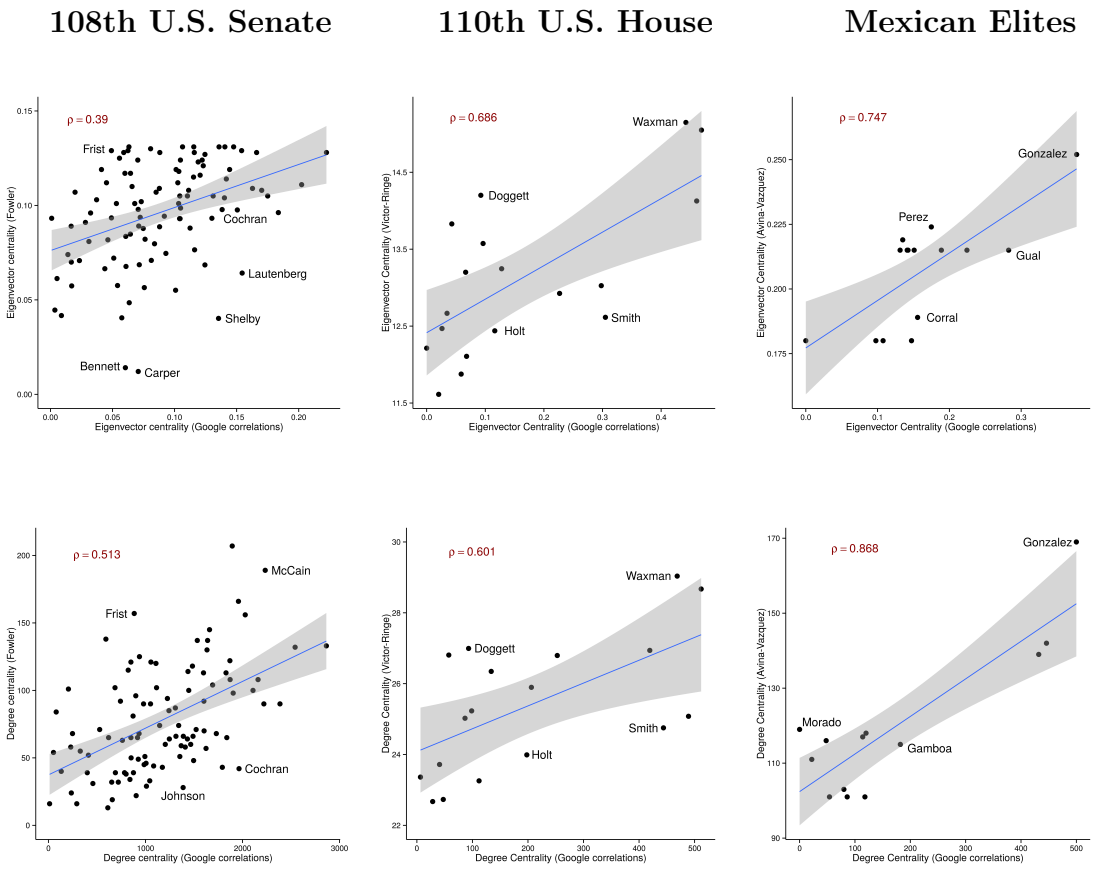


Figure 4.1: Correlations of network centrality (top row: Eigenvector centrality; bottom row: Degree centrality) between data collected using Google correlations versus data collected in previous research.

# CHAPTER 5

# Application: Patronage appointments in Nigeria

In this section, I apply the proposed Google correlations algorithm to address a long-standing question in the comparative study of distributive politics: Do leaders use government appointments as patronage gifts? (Arriola, 2009; Geddes, 2003; Rose-Ackerman, 1999; Stokes et al., 2013)

I test an observable implication of this hypothesis — that leaders use appointments as patronage — by looking at the appointment process to some of the most lucrative government positions: board membership in national oil companies (Victor, Hults and Thurber, 2012). Specifically, I test whether appointments to lucrative state-owned enterprise positions are based on political connections (social connectivity).

I choose Nigeria as a testing ground given it is an extreme case of patronage politics (Sala-i Martin and Subramanian, 2003; Thurber, Emelife and Heller, 2010). If social connections are not linked to government appointments in Nigeria, then it is unlikely to be true elsewhere in the world. Using Nigeria as a case also provides a convenient application for board appointments given that all appointments are typically made concurrently and are made by one government official, the president (as opposed to appointments made by parliaments, cabinets, or councils).

In July 2012, Nigerian President Goodluck Jonathan made eight appointments to the Nigerian National Petroleum Company (NNPC) Board of Directors.[1] Based

[1] See http://fmi.gov.ng/latest/9330/ for details from the Federal Ministry of Information

on previous appointments, the population of potential appointees consists of all 31 executive cabinet ministers and 20 NNPC executive senior officers, making for a network population of $n = 51$.

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

- Restricted dates: one year prior to the board appointments announcement (July 2011 – June 2012);
- Political/social events keywords: "fundraising dinner", "groundbreaking ceremony", "gala", "banquet", "campaign event";
- Restricted web domains to Nigerian newspapers: ngrguardiannews.com, punchng.com, and vanguardngr.com.

Note here that the search is limited to three newspapers in Nigeria. Based on my own informal interviews with Nigerian oil experts,[2] these three papers provide near-comprehensive accounts of all social political events in the country during the specified time period. In this case, it is reasonable to assume that the sample of events identified via the search algorithm is a representative sample of all media-covered, high-profile events in the country. Events not captured by the algorithm can be safely assumed not "high-profile enough" for coverage by one of these three newspapers, and therefore not necessarily of interest when trying to estimate co-occurrence at salient social political events. This is an important point regarding the proposed Google method for network data collection. The sample of media-covered reports searched by the algorithm is very much at the

regarding the July 2012 NNPC board appointments.

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

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

After running the search algorithm, I convert the output to an adjacency matrix — a full list of all non-zero edges between pairs. This matrix can be easily converted to a sociomatrix as described above, as well as visualized in a network graph. In Figure 5.1, I plot the resulting network with pairwise edges (dyads) weighted by co-occurrence of a given pair. Edges that are thicker and darker indicate pairs of individuals that frequently attend the same political events together. The individuals (nodes) are color-coded by appointment status: those in green were not ultimately nominated to the NNPC board, while nodes in blue are the appointees as of July 2012. The red node in the center marks the appointer, President Goodluck Jonathan.

The graph is revealing of several patterns. First, there are few isolated points in the network: all but three individuals (who are not plotted in the graph) are connected to at least one other individual in the network. Second, there is noticeable clustering around the president: almost all nodes are at most two connections away from Goodluck Jonathan. Third, and importantly for the hypothesis of interest, the appointees (in blue) are split into two groups: those who are well-connected within the network and those who are on the periphery of the social network. Indeed, at least two individuals (Diezani Alison-Madueke and Andrew Yakubu) were highly centrally connected prior to being appointed, with four others moderately connected to other members of the network (the remaining two are peripheral).

Network modeling provides a more thorough hypothesis testing exercise. One approach is the Exponential-family Random Graph Model or ERGM (Besag, 1975;

Cranmer and Desmarais, 2011).[3] The ERGM approach offers a model that estimates the effects of node-, dyad- and network-level covariates on social ties taking into account the dependence that is inherent in network data. The general model is 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.

One could use non-network models to analyze these data, such as logistic regression with network attributes as independent variables. However, the conventional regression model ignores selection effects, which is exactly what we are trying to model in this context. That is, appointment to the NOC board is hypothesized to be influenced by an individual's connectivity to others. If being connected to certain individuals tends to influence one's appointment to the board, then assuming separability between the distribution of board appointments and the distribution of the social network is unreasonable (Fellows, 2012).

Using the ERGM specification, I test whether board appointments 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. The latter include network density (edges) and transitivity (edgewise-shared partners). Here, the dependent variable is connectivity — the presence and strength of a tie between two given individuals in the network. I

[3] Other network modeling approaches that have been applied in political science include latent space modeling (Hoff, Raftery and Handcock, 2002; Cao, Prakash and Ward, 2007) and spatial models with endogenous network interdependence (Hays, Kachi and Franzese Jr, 2010).

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 (in July 2012) 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. Though this is a somewhat "roundabout" way of testing the determinants of board appointments — since appointments here are an independent variable — the approach can estimate a correlation between connectivity and appointments, while importantly still accounting for the relational nature of the data.

Results from four model specifications are presented in Table 5.1. In models 1–3, the dependent variable is binary (either a tie exists between two individuals or not) and the specification is the standard ERGM. In model 4, I apply the ERGM-count specification with a discrete count measure of ties. Model diagnostics for the specification including transitivity (column 3) indicate a reasonable fit of the ERGM to the data, with the exception of nodes with 5 to 6 edge-wise shared partners. These graphs are presented in the Appendix.

Across all model specifications, there is a positive correlation between board appointments and social ties. For example, the coefficient for board appointee in model 2 (0.84) implies that there is a 70% probability of a tie[4] forming between two individuals if one is a future board appointee. Compare this to a baseline probability of 50% if ties formed between these individuals by random chance.

One interesting pattern revealed by the network analysis is that there is evidence of clustering. Specifically, the coefficient on the NNPC homophily term

[4]This number is reached by applying the inverse-logit function to the coefficient.

is positive and large in substantive terms, suggesting that if two individuals are both NNPC officials, they are very likely to be connected (between 81% and 95% probability of tie formation based on estimates from models 1 and 3). Indeed, looking at the network graph in Figure 5.1, we can see this clustering of NNPC officials in the nodes to the left of Goodluck Jonathan (in red). Controls for regional clustering, however, do not show evidence that there is grouping by region (province) of origin.

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 0.25 coefficient for board appointees indicates that there is a predicted 28% increase in the count of events co-attended by two individuals if one of them is a board appointee.

While the research design considered here is not strong enough to make causal inferences, these results suggest that being appointed to the NNPC board of directors is positively associated with one's political connectedness (as measured by Google correlations) to others in the network. For the Nigerian case, it seems, there is evidence of the president making patronage-based appointments to the country's lucrative national oil company, as conjectured by previous qualitative scholarly work on NNPC (Nwokeji, 2007; Soares de Oliveira, 2007; Gillies, 2009; Thurber, Emelife and Heller, 2010). Results from network analysis thus serve to complement extant studies by providing systematic evidence for the practice of patronage appointments to Nigeria's state oil company.

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Board appointee | 1.33*** | 0.84*** | 0.69*** | 0.25*** |
| | (0.16) | (0.17) | (0.16) | (0.03) |
| Edges (density) | −3.14*** | −4.81*** | −5.82*** | −0.74*** |
| | (0.20) | (0.32) | (0.46) | (0.11) |
| NNPC homophily | 2.96*** | 2.48*** | 2.24*** | 0.40*** |
| | (0.22) | (0.22) | (0.22) | (0.04) |
| Regional homophily | | −0.17 | −0.18 | |
| | | (0.21) | (0.21) | |
| Pres. Co-ethnic | | 0.18 | 0.16 | |
| | | (0.16) | (0.14) | |
| Google self-hits | | 0.25*** | 0.21*** | 0.87*** |
| | | (0.03) | (0.03) | (0.03) |
| GW-ESP ($\alpha = 0.5$) | | | 0.89*** | |
| | | | (0.27) | |
| AIC | 1206.56 | 1147.91 | 1135.44 | |
| BIC | 1222.12 | 1179.05 | 1171.77 | |
| $N$ (dyads) | 1326 | 1326 | 1326 | 1326 |

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

Table 5.1: Exponential Random-Graph Models (ERGM) of social connectivity and node- and network-level covariates. The table shows results from four model specifications (1–3: ERGM, 4: ERGM-count). Note that coefficients from the first three models are interpreted much the same as logistic coefficients, while those in the fourth model can be interpreted as Poisson regression coefficients.

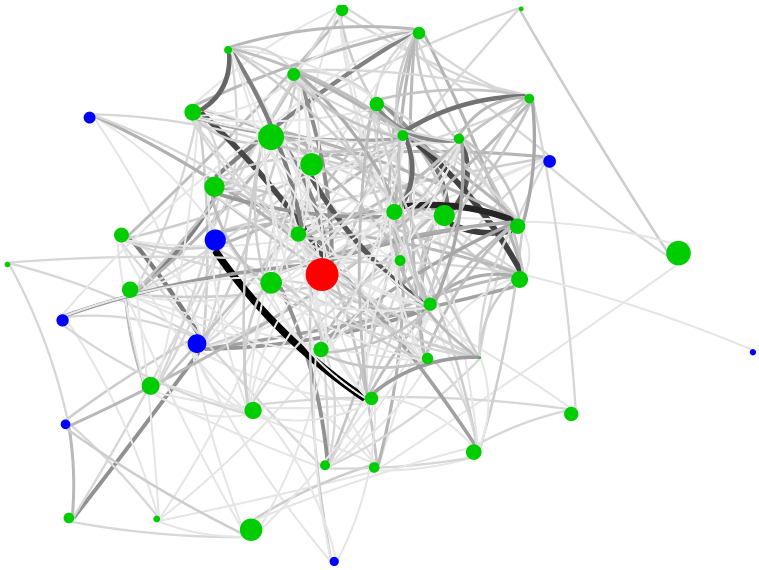

Figure 5.1: Network graph for Nigerian potential appointees to NNPC board of directors. Individuals who were appointed to the board (blue), individuals who were not appointed (green), President Goodluck Jonathan (red).

# CHAPTER 6

# Conclusion

The Google custom search algorithm proposed here provides new opportunities to collecting network data in hard-to-get contexts. This ranges from network 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. Machine learning approaches to network data collection that are currently employed in political science can help address these concerns in contexts where a well-defined corpus of texts for analysis exists, such as in network analysis of legislators, judges, or party members. Beyond these studies, current approaches are limited in their effectiveness in collecting relational data. I have provided one application of the method in the context of patronage networks within Nigerian government appointments to state-owned enterprises. Using the Google search algorithm to assemble network data on Nigerian oil elites, evidence suggests NNPC board appointments are indeed made in part on the basis of political connectivity.

The method is conceptually accurate as cross-validated with three existing political networks — the 108th U.S. Senate (Fowler, 2006), the 110th U.S. House of Representatives (Victor and Ringe, 2009), and Mexican board members in 2012 (Avina-Vazquez and Uddin, 2013) — for which data have been collected using conventional approaches. Despite a reasonable level of conceptual accuracy based on these cross-validations, the proposed method can nonetheless suffer from measurement and sampling errors. First, there is the question of how 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 closely individuals are tied in terms of friendship, ideology, or professional collaboration. Indeed, all measures used in social network analysis suffer from this conceptual problem given that social ties are inherently a latent characteristic that is unobservable by the researcher. More and more cross-validation of the proposed method with existing network data with different measures of social ties can help address this concern. Future research can compare how well measuring ties in terms of event co-occurrence matches up with measures based on self-identified friendship ties, ties inferred from direct observation, and ties inferred from respondent-driven approaches.

Second, there is the issue of sampling error. The Google search algorithm by default is limited to sources that are published in searchable online web pages. Information on event co-occurrence that is published in non-online sources is therefore omitted when constructing network data using this approach. Further, restricting domain names during iterative searches can make the resulting sample of event co-occurrence not representative of the underlying population of social and political events. Ultimately, the algorithm will produce either a full census of results, a random but representative sample of results, or a random, non-representative sample. With unlimited resources, these problems could be addressed by fully digitizing non-online sources and automating the validation of sampled webpages with minimal human-assisted refinement of the algorithm. Future research can provide alternative improvements to the algorithm to reduce this kind of sampling error.

# Appendix

## ERGM Diagnostics

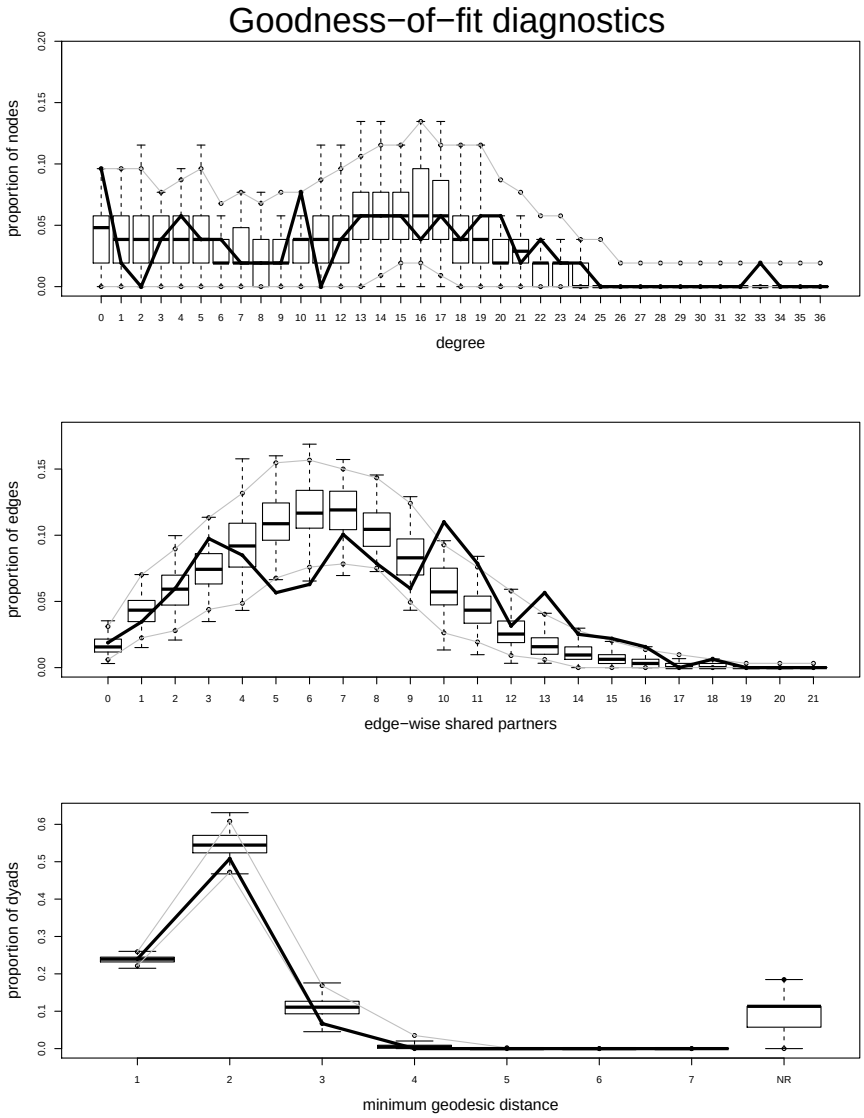


Figure 6.1: Model diagnostics. The graphs visualize ERGM diagnostics for model 3.

## Perl Script for Google Custom Search API

```
#!/usr/local/bin/perl
# written by Paasha Mahdavi 2014 (paasha@ucla.edu)
use strict;
use warnings;
use LWP::UserAgent;
use URI::Escape;
use JSON;
my%configuration = (
        id => '', #Google search API id goes here
        key => '', #Google search API key goes here
        input => '', #input two-column csv file with full list of n*(n-1)/2 pairs in network population
        output => '', #output file name
        expression => '%s+and+%s+SEARCHCRITERIA+site:DOMAIN.COM', #search criteria and restricted domain name(s)
        sleep => 1
);
my$agent = LWP::UserAgent->new;
open( my$input, '<', $configuration{input} ) or die $configuration{input}.': '.$!;
open( my$output, '>', $configuration{output} ) or die $configuration{output}.': '.$!;
while( <$input> ){
        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: '.$value[1]."\n";
        my$response = $agent->get(
                'https://www.googleapis.com/customsearch/v1?'.
                'cx='.$configuration{id}.'&'.
                'key='.$configuration{key}.'&'.
                'q='.sprintf( $configuration{expression}, uri_escape( $value[0] ), uri_escape( $value[1] ) ).'&'.
                'fields=searchInformation/totalResults'
        );
        if( $response->is_success ){
                my$data = decode_json( $response->decoded_content );
                my$hits = $$data{searchInformation}{totalResults};
                print 'response: '.$hits.' result';
                $hits == 1 ? print "\n" : print "s\n";
                print $output ','.$hits."\n";
        }else{
                print 'response: '.$response->status_line."\n";
                print $output ','."\n";
        }
        sleep $configuration{sleep};
}
close $input;
close $output;
print "done\n";
exit;
```

## Python script for Google searching via proxies

```
import csv
import requests
from lxml import html
import random

inputFileName = "termsmexico.csv"
outputFileName = "outputmexico.csv"
rowCounter = 0
rowReport = 1

with open(inputFileName, "rb") as readfile:
    reader = csv.reader(readfile)
    with open(outputFileName,"wb") as writefile:
        writer = csv.writer(writefile)
        currentRow = ["Search Term","Results"]
        writer.writerow(currentRow)

        for row in reader:
            try:
                randInt = random.randint(1,7)
                url = "https://" + str(randInt) +
                      ".PROXYSITE.com/includes/process.php?action=update&u=www.google.com/search?q=" + row[0]
                page = requests.get(url)
                tree = html.fromstring(page.text)
                resultCount = tree.xpath('//div[@id="resultStats"]/text()')
                if len(resultCount) > 0:
                    results = resultCount[0]
                else:
                    results = "No results found"
            except:
                results = "Exception occured"
            currentRow = [row[0] , results]
            writer.writerow(currentRow)
            rowCounter += 1
            if(rowReport > 0):
                if ((rowCounter % rowReport) == 0):
                    print str(rowCounter) + " records processed"
        writefile.close()
    readfile.close()
```

# Bibliography


Alemán, Eduardo and Ernesto Calvo. 2013. "Explaining policy ties in presidential congresses: A network analysis of bill initiation data." *Political Studies* 61(2):356–377.

Arriola, Leonardo R. 2009. "Patronage and Political Stability in Africa." *Comparative Political Studies* 42:1339–1362.

Avina-Vazquez, Carlos Rafael and Shahzad Uddin. 2013. "Network of Board of Directors in Mexican Corporations: A Social Network Analysis." APIRA Conference Proceedings.

Berthon, Pierre, Leyland Pitt and Gerard Prendergast. 1997. "Visits, Hits, Caching and Counting on the World Wide Web: Old Wine in New Bottles?" *Internet Research* 7(1):5–8.

Besag, Julian. 1975. "Statistical analysis of non-lattice data." *Journal of the Royal Statistical Society. Series D (The Statistician)* 24:179–195.

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle and James H Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature* 489(7415):295–298.

Box-Steffensmeier, Janet M. and Dino P. Christenson. 2014. "The evolution and formation of amicus curiae networks." *Social Networks* 36(0):82 – 96. Special Issue on Political Networks.

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

Cao, Xun, Aseem Prakash and Michael D Ward. 2007. "Protecting jobs in the age of globalization: examining the relative salience of social welfare and industrial subsidies in OECD countries." *International Studies Quarterly* 51(2):301–327.

Carpenter, Mason A and James 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.

Chadefaux, Thomas. 2014. "Early warning signal for war in the news." *Journal of Peace Research* 51(1):5–18.

Clifton, Brian. 2012. *Advanced Web Metrics with Google Analytics*. Wiley.

Cohen, Marty, David Karol, Hans Noel and John Zaller. 2009. *The Party Decides: Presidential Nominations Before and After Reform*. Chicago: University of Chicago Press.

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

Davis, J.A., B. Gardner and M.R. Gardner. 1941. *Deep South*. Chicago: University of Chicago Press.

Fellows, Ian. 2012. Exponential Family Random Network Models PhD thesis University of California, Los Angeles.

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

Fowler, James H., Timothy R. Johnson, James F. Spriggs, Sangick Jeon and Paul J. Wahlbeck. 2007. "Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court." *Political Analysis* 15(3):324–346.

Galaskiewicz, J. 1985. *Social Organization of an Urban Grants Economy.* New York: Academic Press.

Geddes, Barbara. 2003. *Paradigms and Sand Castles.* Cambridge University Press.

Gile, Krista J and Mark S Handcock. 2010. "Respondent-Driven Sampling: An Assessment of Current Methodology." *Sociological methodology* 40(1):285–327.

Gillies, Alexandra. 2009. "Reforming corruption out of Nigerian oil?" *Chr. Michelson Institute U4 Brief* 2.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* 21:267–97.

Grossman, Guy. 2014. "Do Selection Rules Affect Leader Responsiveness? Evidence from Rural Uganda." *Quarterly Journal of Political Science* 9(1):1–44.

Hafner-Burton, Emilie M, Miles Kahler and Alexander H Montgomery. 2009. "Network analysis for international relations." *International Organization* 63(03):559–592.

Hays, Jude C, Aya Kachi and Robert J Franzese Jr. 2010. "A spatial model incorporating dynamic, endogenous network interdependence: A political science application." *Statistical Methodology* 7(3):406–428.

Hoff, Peter D, Adrian E Raftery and Mark S Handcock. 2002. "Latent space approaches to social network analysis." *Journal of the American Statistical Association* 97(460):1090–1098.

Holland, Paul W and Samuel Leinhardt. 1973. "The structural implications of measurement error in sociometry†." *Journal of Mathematical Sociology* 3(1):85–111.

Jones, Jason J., Jaime E. Settle, Robert M. Bond, Christopher J. Fariss, Cameron Marlow and James H. Fowler. 2013. "Inferring Tie Strength from Online Directed Behavior." *PLoS ONE* 8(1):e52168.

Kim, Jang Hyun, George A. Barnett and Han Woo Park. 2010. "A Hyperlink and Issue Network Analysis of the United States Senate: A Rediscovery of the Web as a Relational and Topical Medium." *Journal of the American Society for Information Science and Technology* 61(8):1598–1611.

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

Koger, Gregory, Seth Masket and Hans Noel. 2010. "Cooperative party factions in American politics." *American Politics Research* 38(1):33–53.

LaCour, Michael J. and Donald Green. 2014. "Messages, Messengers, and Diffusion of Support for Gay Equality: Results from Two Longitudinal Field Experiments." UCLA manuscript.

Lange, Margaret M. 2014. A Comparison of Estimators for Respondent-Driven Sampling M.s. dissertation UCLA.

Lee, Sang Hoon, Pan-Jun Kim, Yong-Yeol Ahn and Hawoong Jeong. 2010. "Googling Social Interactions: Web Search Engine Based Social Network Construction." *PLoS ONE* 5(7):e11233.

Lupu, Yonatan and Erik Voeten. 2012. "Precedent in international courts: A network analysis of case citations by the European court of human rights." *British Journal of Political Science* 42(02):413–439.

Moore, Gwen. 1979. "The Structure of a National Elite Network." *American Sociological Review* 44(5):pp. 673–692.

Morgan, Jason. 2014. "The Latent Path Model for Dynamic Networks." Ohio State University Manuscript.

Nwokeji, G. U. 2007. The Nigerian National Petroleum Corporation and the Development of the Nigerian Oil and Gas Industry: History, Strategies and Current Directions. Technical report Houston, James A. Baker III Institute for Public Policy of Rice University.

Rose-Ackerman, Susan. 1999. *Corruption and Government: Causes, Consequences, and Reform*. New York: Cambridge University Press.

Sala-i Martin, Xavier and Arvind Subramanian. 2003. "Addressing the Natural Resource Curse: An Illustration from Nigeria." *NBER Working Paper* .

Sampson, Samuel F. 1969. Crisis in a cloister: A sociological analysis of social relationships and change in a novitiate PhD thesis Cornell University.

Singerman, Diane. 1995. *Avenues of Participation: Family, Politics, and Networks in Urban Quarters of Cairo*. Princeton, NJ: Princeton Univ. Press.

Soares de Oliveira, Ricardo. 2007. *Oil and Politics in the Gulf of Guinea*. London: Hurst and Company.

Stokes, Susan C, Thad Dunning, Marcelo Nazareno and Valeria Brusco. 2013. *Brokers, Voters, and Clientelism: the puzzle of distributive politics*. Cambridge University Press.

Thurber, Mark C., Ify Emelife and Patrick R.P. Heller. 2010. "NNPC and Nigeria's Oil Patronage Ecosystem." *PESD Working Paper* 95.

Victor, David G., David Hults and Mark C. Thurber, eds. 2012. *Oil and Governance: State-owned Enterprises and the World Energy Supply*. Cambridge, UK: Cambridge University Press.

Victor, Jennifer Nicoll and Nils 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.

Wasserman, Stanley and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.

White, Douglas and Ulla Johansen. 2005. *Network analysis and ethnographic problems: Process models of a Turkish nomad clan*. Lexington books.