BRILL

ASIAN JOURNAL OF
SOCIAL SCIENCE 43 (2015) 588–612

Asian Journal
of Social Science
brill.com/ajss

Elite Co-Occurrence in the Media*

Vincent A. Traag
KITLV/Royal Netherlands Institute of Southeast Asian and Caribbean Studies

Ridho Reinanda
ISLA, University of Amsterdam, The Netherlands

Gerry van Klinken
KITLV/Royal Netherlands Institute of Southeast Asian and Caribbean Studies

Abstract

We present a new computational methodology to identify national political elites, and demonstrate it for Indonesia. On the basis that elites have an "organised capacity to make real and continuing political trouble", we identify them as those individuals who occur most frequently in a large corpus of politically-oriented newspaper articles. Doing this requires mainly well-established named entity recognition techniques and appears to work well. More ambitiously, we also experiment with a new technique to map the relational networks among them. To establish these networks, we assume that individuals co-occurring in one sentence are related. The co-occurrence technique has rarely been applied to identify elite networks. The resulting network has a core-periphery structure. Although this in line with our sociological expectations of an elite network, we find that this structure does not differ significantly from that of a randomly generated co-occurrence network. We explain that this unexpected result arises as an artefact of the data. Finally, we assess the future potential of our elite network mapping technique. We conclude it remains promising, but only if we are able to add more sociological meaning to relations between elites.

* All authors are researchers with the Elite Network Shifts project. This research is funded by the Royal Netherlands Academy of Arts and Sciences (KNAW) through its e-Humanities project. http://www.ehumanities.nl/computational-humanities/elite-network-shifts/.

 | DOI: 10.1163/15685314-04305005

Keywords

co-occurrence network – media network – elites

Introduction

National elites play pivotal roles in political, economical, cultural and intellectual life (Mills, 2000; Putnam, 1976). This is not to argue that the elite will always behave as a singe entity. On the contrary, the elite may be rife with competition and conflict, and such divisions in the elite are often important in economic developments (Brezis and Temin, 2008) or societal upheaval (Turchin, 2005), such as protests (Burton, 1984), revolutions (Goldstone, 1991) and war (Eff and Routon, 2012). But the elite itself is also shaped by such events. If a revolution has succeeded, leaders from the old regime might be ousted, and replaced by new revolutionary cadres, thus producing elite rotation (Dogan and Higley, 1998).

How elites are connected to each other may thus have a clear political impact. Totalitarian regimes appear to have highly unified national elites. A moderately differentiated elite is characteristic of a pluralist democracy. A deeply divided elite may produce political crisis and even violent factionalism (Goldstone, 2001). Knowing the structure of elite networks may help in understanding or explaining larger societal developments.

The research challenge we address consists of two steps. The methodology we use for both tasks differs from traditional approaches in the social sciences. First, we identify a list of national elites. To solve this well-known and fundamental problem in elite studies, we offer a new, digital method. Second, we automatically calculate the networks by which these elites are connected with each other. This problem is fundamental to elite studies, but before the use of digital methods it has not been possible to solve on a national scale.

To identify national elites, social scientists previously used a variety of sources to manually assess whether somebody is an elite, for example, based on their formal position or their role in community politics. We take a rather different approach, and base our analysis on people who appear regularly in the newspaper. Following Higley and Burton (2006) we define the elite as people who have the "organised capacity to make real and continuing political trouble". Surely, not all the people who appear in the newspaper are elites according to this definition, nor do all elites appear in the newspaper (whichever definition of elites we use). Nonetheless, people who regularly appear in the (political) news do so for a reason. In one way or another, journalists took an interest

in those people, and decided to use their name in a number of articles. This signals that at least such people are of some public interest, and as such may be considered to have the capacity to make real and continuing political trouble. We are interested in Indonesia in particular. In the highly personalised politics of this Southeast Asian nation, this approach could represent a particularly promising route towards identifying national elites both formal and informal.

Our second, and central, question is how the Indonesian elite connect to each other. Do well connected elites connect to each other? Or do they tend to be linked to less well connected elite? Are weak links—infrequent co-occurrences—necessary to keep the network connected, as in the "strength of weak ties" hypothesis (Granovetter, 1973)? This would suggest that most people tend to appear in clusters, so that they frequently occur with each other, and only sometimes with people outside of their cluster. Is the network a small world, reminiscent of the famous "six degrees of separation" (Travers and Milgram, 1969)? This would mean that even if two people are not immediately connected, only few intermediaries might be necessary to connect them.

We base our analysis on about 140,000 newspaper articles from a news service called *Joyo*.[1] The *Joyo* corpus consists of manually-selected English-language articles downloaded from the websites of prominent newspapers and emailed to a list of subscribers interested mainly in the politics and economics of Indonesia. Our corpus consists of articles from about 2004–2012. The largest number (nearly 50,000) derive from *The Jakarta Post*. This liberal daily published in the Indonesian capital is read by expatriates and upper middle-class Indonesians. Articles also came from newspapers published in Singapore, Malaysia, Hong Kong, Australia, the US and wireline agencies.

To begin with the question of the identification of national elites, we developed an automated identification process that moreover introduces very few arbitrary assumptions. The corpus is simply too large for manual coding. By using automated coding, we are able to process the complete corpus, and provide a relatively complete picture of who appears in this corpus. This automation and large dataset allows a bird's-eye view of how the elite is portrayed in the media. We checked a small percentage of the resulting names. Although some of the extracted names also include some errors, such as place names, organisations and single (first) names, most of the found entities refer to persons. This is especially of interest for a developing country, for which data is usually much more difficult to obtain.

1 http://www.joyonews.org.

Moving on to the elite network question, we first give a concise introduction to some network terminology. After providing some details on the data and the methodology, we discuss our results. Our central question is rather exploratory: Does the automated technique we deploy have the power to yield sociologically meaningful results? We test this by first of all examining the structure of the resulting network—does it look reasonable, given what we know about these elites from other studies? Next, we ask what mechanisms may lie behind the observed network—can we exclude all but sociological mechanisms for its origin, or could it also arise from mechanisms extraneous to social reality? Here, the answers turns out to be surprising.

Network

We will first briefly introduce some essential network concepts. We will use a minimum of formulas, so as to as remain as tangible and intelligible as possible for a broader audience. Nonetheless, some subtleties might be lost in translation. We will make clear how we constructed the network. Finally, we will discuss our results.[2]

Introduction

Complex networks have been a prominent research topic for the past decade. One of the reasons is that complex networks appear in a multitude of scientific disciplines, varying from neurology (Bullmore and Sporns, 2009; Hagmann et al., 2008), ecology (Garlaschelli et al., 2003; Guimerà et al., 2010) to international relations (Cranmer et al., 2014; Maoz et al., 2008; Garlaschelli and Loffredo, 2005) and human mobility (González et al., 2008; Simini et al., 2012) providing a unified theoretical framework for analysis. Although many characteristics of networks, such as a broad distribution of the number of links, seem to be (nearly) universal (Barabási, 2009), there are also some noteworthy differences between different types of networks (Petri et al., 2013; Guimerà et al., 2007; Amaral et al., 2000). For an introduction in social networks for social scientists, see Wasserman and Faust (1994), while Newman (2010) provides a more mathematical introduction into complex networks.

Formally speaking, a network is just a collection of points and a collection of lines between those points. The points are usually called *nodes* or *vertices*, and the lines between them are called *links*, *edges*, *arcs* or *ties* (these words do

2 A more technical version of this paper is also available (Traag et al., 2014).

not imply any difference in this case and we might use these different words interchangeably). In mathematics, a network is sometimes called a *graph*. We usually simply number the nodes 1, 2, …, n, where n thus denotes the number of nodes in the network. Edges are then simply represented as a pair of nodes: For example, (5, 10) denotes that there is an edge between node 5 and node 10. In this case, we do not care about the direction of an edge, so that edge (5, 10) refers to exactly the same edge as would (10, 5). We sometimes say that a link connects two endpoints, so that for link (5, 10) both node 5 and node 10 are endpoints. We only treat here undirected networks, and some of the concepts may work differently in directed networks. What the links and nodes represent is another question, and this depends on more substantive concerns.

Let us briefly look how many links a network can have. Letus look at a network of 10 nodes. Let us say we start counting from node 1. It can then have a link to node 2, 3, …, 10, so that is 9 different possibilities. Now for node 2, it can connect to node 1, 3, 4, …, 10, so again 9 possibilities. But then of course we count the link (1, 2) twice. If we continue to do so for all nodes, we will have counted each link twice. So, the total number of possible links in a network of 10 nodes is $\frac{10\times 9}{2} = 45$. In general then, the total number of possible links in a graph with n nodes is $\binom{n}{2} = \frac{n(n-1)}{2}$

The actual number of links divided by the number of possible links is called the *density*. If the density of a graph is 0, this means that there is not a single link present. If the density of a graph is 1, this means that all possible edges are present. In most real networks, the density tends to be rather low, since it is usually impossible for a node to have links with many of the other nodes. If you think of friends, for example, people tend to have only a couple of real good friends, a couple of dozen more distant friends, and perhaps a few hundred acquaintances. So, if the network covers a million people or so, only a tiny proportion of the actual number of links will exist.

We call the set of nodes that are connected to some node i the *neighbours* of node i. The number of neighbours of node i is called the *degree* of node i. Hence, if a node has 7 neighbours, then it has a degree of 7. Notice that since each edge connects two nodes, they are each other's neighbours. Compared to the density, it often makes much more sense to look at the average degree: How many links does a node have on average? This tends to depend less on the size of the network. As in the previous example, let's say that on average people tend to have about 10 friends. Then, this should stay about the same, regardless of whether we look at a network of a thousand people or a million people.

Even though not everybody is immediately connected to everybody, it might be that there is some *path* between them. So, if node 1 is not connected to

node 3, but both are connected to node 2, we may take a path from node 1 to node 2 and then to node 3, tracing only existing edges. If there exists such a path from everybody to everybody, the network is said to be *connected*. Sometimes though, some nodes are unconnected to the rest of the network. Although those nodes might be connected via some path among themselves, they are not connected to the rest. We can thus divide the network into several *connected components*. Everybody within a connected components is able to reach everybody else within that same component, while it is impossible to reach somebody in another connected component. Almost always most real networks consist of one very large connected component and possibly many other small connected components consisting of only a few nodes (Newman, 2010).

In our case, the network will be weighted. For weighted graphs, each edge has a (positive) weight associated to it. If there is no edge between two nodes, there is also no weight. Now instead of counting the number of neighbours of node *i*, we can sum the weight of all the edges from node *i*, which is called the *strength*. Hence, the strength can be viewed as a sort of weighted degree. If the weight would be 1 for each edge, the strength and the degree would be exactly equal. We commonly refer to links that have a low weight as *weak links*, while *strong links* refer to links that have a high weight.

Finally, one concept that is particularly relevant in social networks is the *clustering coefficient*. For a certain node *i* this is equal to the proportion of connections amongst the neighbours of *i*. For example, if node *i* has 5 neighbours, there are $\frac{5\times4}{2} = 10$ possible connections among the neighbours. If then 6 of these links exist, the clustering coefficient is $\frac{6}{10} = 0.6$. Often in social networks, people that are friends with somebody else, also tend to be friends themselves, which will lead to a relatively high clustering coefficient (Watts and Strogatz, 1998).

A related concept, but then from the perspective of a link is the *overlap*, which is the number of common neighbours. Consider some link (3, 7), and that node 3 is connected to nodes 1, 4 and 7, and that node 7 is connected to nodes 3, 4 and 5. Then, nodes 3 and 7 have only one node in common (node 4), while in total they connect to three nodes besides themselves (i.e., nodes 1, 4 and 5). So, the overlap is $1/3 \approx 0.33$. A higher overlap means that people tend to have neighbours in common, similar to the idea of the clustering coefficient.

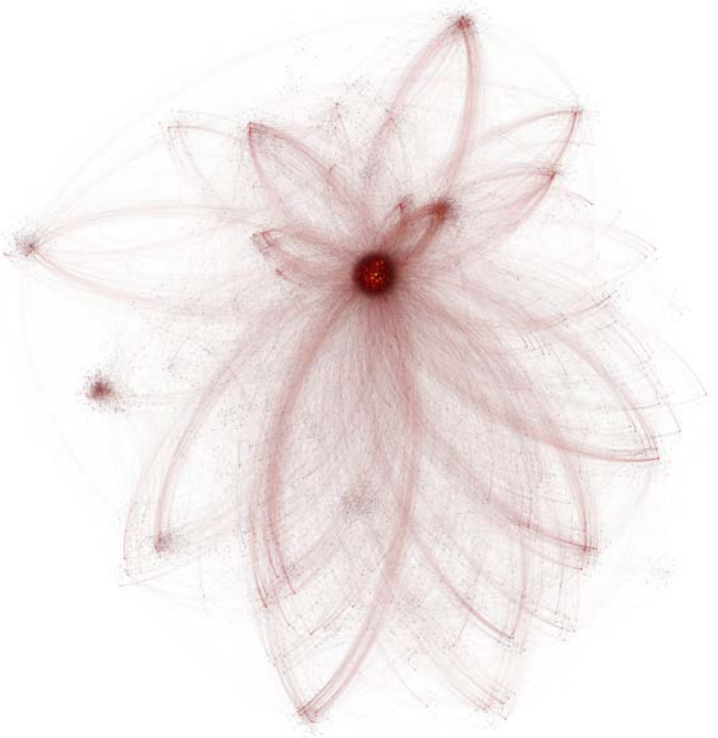

FIGURE 1 *A visualisation of the network with 9,567 nodes. The average degree is about 12 and the average weight of a link about 3. The size of the nodes in the visualisation is proportional to the degree. The width of the links is proportional to the weight. This visualisation is produced using the OpenOrd layout algorithm in Gephi.*

Constructing the Network

We base our analysis on a data set consisting of 140,263 articles. The *Joyo*[3] data set, as explained above, covers politically relevant Indonesian news in English from roughly 2004 to 2012. Sports and entertainment is less covered. Since we are interested in political elites, this eases our task somewhat.

The first task is to automatically extract names of persons from the text. This can be automatically done by a technique[4] known as *named entity recognition* (Finkel et al., 2005). This technique scans the text and automatically tags parts of sentences as being "entities", and classifies them into a number of distinct

3 http://www.joyonews.org.

4 http://nlp.stanford.edu/software/CRF-NER.shtml.

categories (e.g., organisation, locations, persons). Such techniques are not infallible and mistakes will be made, but at a relatively low rate of around 10–15% (Finkel et al., 2005). We should not forget though that human efforts will also contain considerable coding errors.

We only include persons (i.e., entities that are classified as "person" by the named entity recognition) in our network, and only people that occurred in more articles than average. Many of the people that appear in our network might be regarded as being part of the elite to some extent. However, not all people necessarily come from Indonesia, as also other foreign news is included that is relevant to Indonesia. For example, elites from Malaysia, the Philippines and the US also regularly feature in these articles. Although these people might not be part of a domestic Indonesian elite, they surely may have some influence in domestic developments, and as such may be relevant.

Once all persons have been identified, we have to disambiguate them. There are generally two types of errors that can be made with names (Milne and Witten, 2008): (1) a single name corresponds to two different persons (e.g., "Bush" can refer to the 43rd or 41st US president); and (2) two different names refer to the same person ("President Clinton" or "Bill Clinton" both refer to the 42nd US president). The second problem appears much more prominent than the first problem in our corpus, as people are generally referred to in many different ways in journalistic prose (including or not positions, titles, initials, maiden names, etc.).

We disambiguated these names by using a combination of similarity measures based on Wikipedia matching, name similarity and network similarity. The more prominent people often have a Wikipedia page, and various spelling variants are redirected to the same entity (i.e. "President Clinton" and "Bill Clinton" both redirect to the same Wikipedia page). We find there are 5,619 names that have a match in either an English or Indonesian Wikipedia. Name similarity is simply defined as how similar the two different names are using a type of edit distance measure. Network similarity is defined as the proportion of contacts two people have in common—if two people are, in fact, the same, we may expect them to occur with similar people. We construct an average similarity between all entities, which we then cluster such that each cluster should have an average internal similarity of 0.85 (all similarity measures are between 0 and 1), using a technique called the *Constant Potts Model* (Traag et al., 2011). Indeed, using this technique, Hicks et al. (2015) shows we can identify elites.

Now, we come to the second task, namely to construct a network linking those we have identified as prominent in the news. We derive the links also from the text itself. To do this, we make a very simple assumption, namely

that persons mentioned together have some kind of relationship. This is not always true, but our tests have shown that it works acceptably well as a first approximation (Reinanda et al., 2013).

We look at whether people appear in the same sentence. The appearance of people in sentences can also be represented as a network, a so-called *bipartite* network. In such a bipartite network, we create two different kinds of nodes: people and sentences. Whenever a person occurs in a sentence, we create a link between that person and that sentence. Notice that such a link represents the idea of "occurs in", and all links will always be between persons and sentences, never between persons themselves, or between sentences themselves. This is called a *bipartite structure*, and it will play an important role in this paper.

Since we are interested in people, rather than the bipartite structure of people occurring in sentences, we look at co-occurrence. This means that we will create a link whenever two people occur in the same sentence. We only take into account the unique names in a sentence, so that if the same name is mentioned multiple times in the same sentence, this does not add an additional co-occurrence (although this is quite rare). Of course, more than two people might occur in the same sentence. The co-occurrences for a given sentence are then simply all possible combinations of all mentioned (unique) names. So, if three people (1, 2 and 3) co-occur in the same sentence, this corresponds to three links: (1, 2), (1, 3) and (2, 3). We do this for every sentence, and count in how many sentences such a co-occurrence was observed. The number of sentences in which two people co-occur then constitutes the weight of that link. This network consisting only of people thus derives from the bipartite network of people and sentences. The network of people is connected and constitutes the network we analyse in this paper. A visualisation of this network is displayed in Figure 1.

Of course, what co-occurrence exactly implies is not always clear: two people might be mentioned together, for example, because they collaborate, or because they are contestants in an election. Hence, we cannot say if two people that co-occur have any more significant relationship: do they know each other? Have they ever communicated? Have they met face to face? Are they close friends? Sworn enemies? We simply cannot tell (yet). This is essential to bear in mind when drawing any conclusions: The network is based on co-occurrence, not on "actual relationships".

Nonetheless, a co-occurrence signifies something (although we cannot say what exactly). For some reason, a journalist decided to name them together in a single sentence. Even though a co-occurrence might not coincide with any one single definition of a "relationship" in the sociological sense (Knoke and Kuklinski, 1982), it may reveal something about how people are connected in

the media. Moreover, it tells us something about how certain relationship are perceived, and paraphrasing Thomas and Thomas (1928), if they are perceived as real, they will have real consequences. As such, we may hope that extracting relationships at a large scale will reveal how members of the elite in Indonesia connect with one another. Given the possibilities of using co-occurrences in newspaper articles for constructing networks, there have been surprisingly few earlier studies of such networks (Steinberger and Pouliquen, 2007; Pouliquen et al., 2008; Özgür and Bingol, 2004; Joshi and Gatica-Perez, 2006).

Results

Empirical Results

Let us now examine the visualisation of the resulting network in the figure. What does it look like? What accounts for its structure? What, if anything, does it tell us about Indonesian elite relations?

The elite co-occurrence network has 9,567 nodes and 59,182 edges. This means that, on average, people have about 12 neighbours. In total, we recorded 174,374 co-occurrence between people in this network. So, on average, somebody co-occurs about three times with somebody else, and in total will co-occur with its neighbours about 36 times. As we will see, however, many nodes are very different from this average scenario.

While most of the nodes have only one or two neighbours, there are a few nodes that have many neighbours, which we call *hubs*, some even reaching 2,000 neighbours. For those acquainted with Indonesian politics, the names of these hubs should be familiar. The first hubs are Yudhoyono (president, 2004–2014), Suharto (president, 1967–1998), Megawati (president, 2001–2004) and Kalla (vice-president, 2004–2009). However, it also includes people from outside Indonesia, such as Obama and Bush. Near the bottom of the list we find people such as Wiyogo Atmodarminto, a retired army general and former governor of Jakarta, Iwan Piliang, a journalist and activist, and Hendi Prio Santoso, President Director of the largest gas company of Indonesia, to name but a few. To give an impression of the enormous difference, more than 70% of the people have less than ten links, while only seven nodes have more than 500 links (Figure 2). Although this might come as a surprise to some, this is found in many empirical networks (Barabási and Albert, 1999). The few prominent people thus draw most of the attention, whereas most others remain more obscure.

The weight itself is also quite broadly distributed. Similar to the degree, most of the links have a low weight of only one or two, while only eight links have a

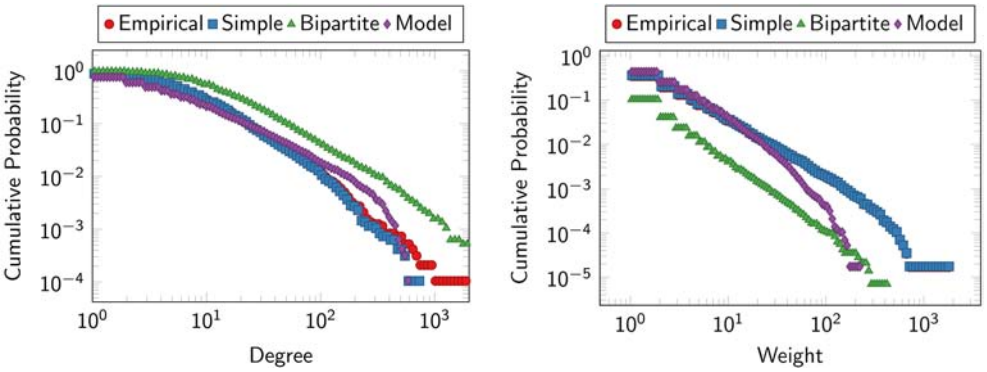

FIGURE 2 *Distributions*

weight of over 500 (Figure 2). The most frequently co-occurring pair of people are Yudhoyono and Megawati with a frequency of 2,187. In addition to this large heterogeneity, we find that people that have more links have a higher average weight per link (Figure 3). To illustrate, Megawati has a degree of 735 and an average weight per link of about 12, while Hendi Prio Santoso has a degree of three and an average weight of one. This means that the hubs do not only occur with many more people, they also co-occur more frequently with them.

Everybody can reach everybody in this network within ten steps, and almost 98% of the time, two people can reach each other within five steps. On average, you need only 3.46 steps to reach somebody else. For example, although the two presidential candidates of the 2014 elections, Joko Widodo and Prabowo Subianto, did not (yet) co-occur in our corpus, they are connected because both co-occurred with Mohammad Hatta, Indonesia's first vice-president, so are at a distance of two of each other. So, everybody is connected to everybody else in relatively few steps. Again, this is no surprise for those familiar with complex networks, as it is the case in most complex networks, know as the famous "six degrees of separation" (Watts and Strogatz, 1998). In part, this is also due to the presence of hubs, many of the shortest paths go via a hub, similar as you would when travelling: you take a bus or train to the airport (a hub) and from there you make your way to your destination. So, even if people are not immediately connected, they are likely to be connected through a common third party.

On average, relatively many neighbours of a person tend to be connected amongst themselves. The average clustering coefficient, as introduced earlier, is about 0.30, so that on average about 30% of the pairs of neighbours of a node are also connected to each other. For example, Joko Widodo has 26

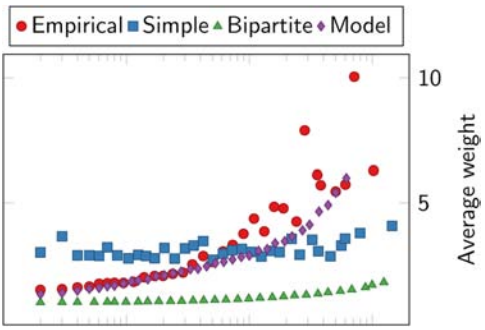

FIGURE 3 *Degree versus average weight*

neighbours, and with 38 links between them, they represent about 12% of the possible number of links amongst Widodo's neighbours. Another example, Hasyim Muzadi, former running mate of Megawati in the 2004 elections, has only five neighbours, but with ten links all of them are connected. This is in fact quite common in this network: Nodes that have a lower degree (i.e., have fewer neighbours) tend to have a higher clustering coefficient. The average clustering coefficient for nodes having three links is about 40%, while for a degree of around 100 this drops to about 11% (Figure 4). This is usually associated with an idea of hierarchy: Nodes that have many links connect many people that are otherwise unconnected, while the few neighbours of nodes with a low degree are also connected amongst themselves. So the hubs tend to connect disparate people that are locally clustered. There is also a weighted variant of this clustering coefficient, which shows that the clustering is slightly biased towards stronger links. Similarly, it still shows that the hubs tend to have lower clustering coefficients.

If these hubs tend to connect disparate parts of the network, we may wonder what types of persons they connect. In particular, do they tend to connect to other hubs, or would they tend to connect to lower degree nodes? We can see that by looking at the average degree of the neighbours of a node. For example, Yudhoyono has an impressive 2,099 neighbours, but they have a degree of only about 30 on average, while, Mohamed Yakcop, a Malaysian minister, co-occurred with only three people, but they have an average degree of 125. This is characteristic for the larger pattern we see: high degree nodes tend to connect to low degree nodes (Figure 5). This is again in line with hubs connecting disparate parts of the network. But if we look at the weighted variant, we see that the average neighbour degree is actually increasing with degree (Figure 5). This means that although the hubs mainly connect to low degree nodes, they seem to do so only through relatively weak links, while they

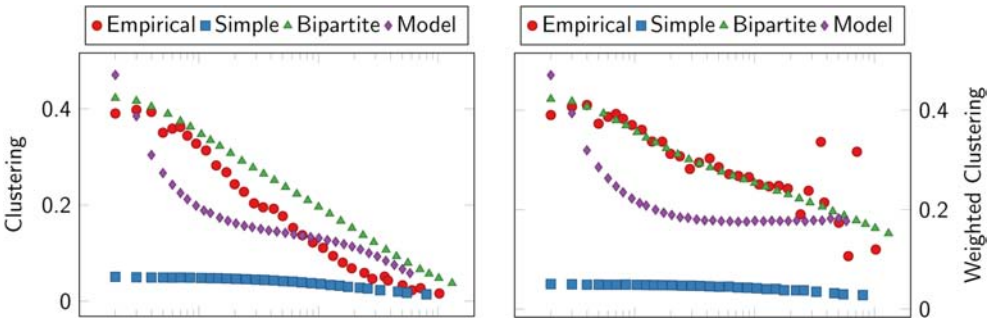

FIGURE 4 *Clustering coefficient*

connect to other hubs with a much higher weight. So, those hubs co-occur with many people that have only a low degree, but they do so only a few times, while they co-occur frequently with other hubs. For example, Andi Mallarangeng, former spokesman for Yudhoyono, has a degree of 199 and an average neighbour degree of 94, but he co-occurs most frequently, about 25% of the time with Yudhoyono (a degree of 2,099); about 10% of the time with Anas Urbaningrum (a degree of 145), chairman of the Democratic Party; and about 5% of the time with Marzuki Alie (a degree of 85), while he co-occurs only once with, for example, Agus Yudhoyono Harimurti, son of Yudhoyono, who has only a degree of 22. So, if a link involves two persons that have many links, they tend to co-occur more frequently.

This points to an interesting hypothesis suggested by Granovetter (Granovetter, 1973): Weak links tend to connect people that do not have many neighbours in common. Strong links are expected to have a high overlap, and weak links to have a low overlap. Indeed, we also see this in our data (Figure 6). Combining this with our previous observation, we can conclude that the weight of a link is mostly determined by two factors: (1) what the degree of the two endpoints of the link is; and (2) how many neighbours these two endpoints have in common. Knowing only these two properties, we get a pretty good idea of what the weight should be.

In summary, we have the following observations:

- **Heterogeneity:** The degree varies enormously, with some hubs and many low degree nodes, as does the weight.
- **Strong hubs:** The hubs have a higher average weight per link.
- **Small world:** It is a small-world network, with the hubs connecting people.
- **Hierarchical hubs:** The hubs are less clustered than the other nodes, slightly biased towards stronger links.
- **Cohesive hubs:** Although hubs mainly connect to low degree nodes, they do

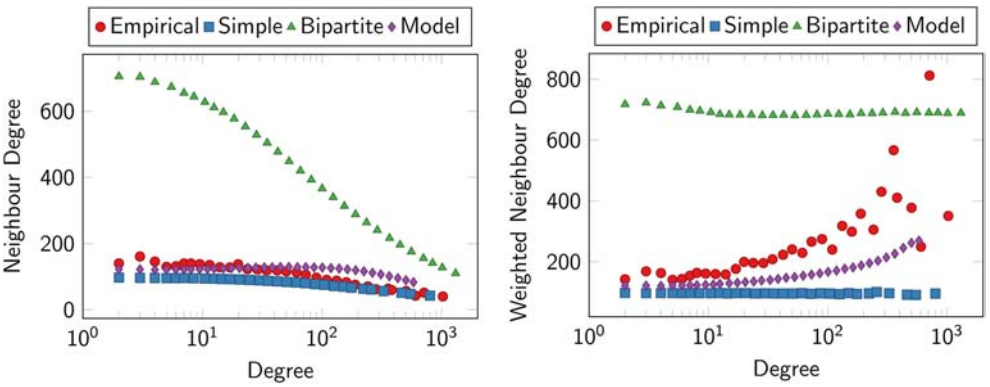

FIGURE 5 *Neighbour degree*

so with a relatively low weight, while they connect much stronger with other hubs.

– **Strong overlapping links:** Stronger links tend to occur between hubs that have many common neighbours.

Putting all this together, we can indeed characterise the structure of our network as a core-periphery structure (Csermely et al., 2013). The core consists of the hubs, while the periphery is made up of lower degree nodes. The core connects mostly to each other through strong links, while the links between the core and the periphery are much weaker. The periphery itself is locally clustered; however, the clusters in the periphery do not connect to the rest of the periphery, but are mostly connected to the core.

This is also confirmed by a cluster analysis (Traag et al., 2011) (known as "community detection") of our network, which mostly shows one big central community (i.e., our core, mostly concerning Indonesian politics) with many more peripheral clusters surrounding the core. These communities usually seem to be related to either some issue or foreign politics. There is, for example, a community that revolves around Malay politics. Another community seems mostly about Thai and Burmese politics, and is presumably also related to a series of incidents along the Thai-Burmese border. Yet another community focuses on East Timor. A fourth community concerns Philippine politics, and is also related to terrorism from the Abu Sayyaf group. Interestingly, there is also a community that consists of journalists. Another community consists surprisingly enough of place names or locations: some of the misses by the named entity recognition. Although most of the corpus is politically oriented, some artists from the film and music industry also appear in a cluster. Finally, various scandals revolve around a few main figures, which often pops up as a community in itself as well.

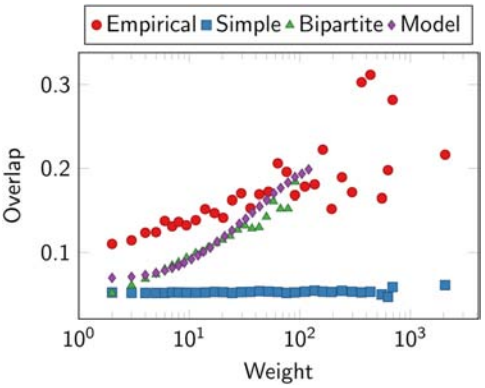

FIGURE 6 *Weight versus overlap*

Randomisation and Model

The network looks sociologically realistic. Elites stand at the peak of a hierarchical society, and they are also stratified within themselves. A core-periphery structure fits with our expectations. We do expect to see a central hub of top elites, such as the president and entourage, who connect mainly with each other, and a periphery made up of local clusters of second-rung elites, who relate mainly to the central hub. That the Granovetter hypothesis appears to be applicable adds to the resemblance between the calculated co-occurrence network and some social reality. However, resemblance is not the same as explanation.

To study the question of mechanisms, we will have to proceed with due caution. We will look at random networks (sometimes called random graphs) that adhere to the constraints provided by the empirical network. These constraints are in this case: the number of nodes and edges, the number of links each node has, and the weight that a certain link has. We can then create a random graph that adheres to these constraints. By analysing how likely our results from the empirical network are, compared to this random graph, we will know whether some of these properties stem from these constraints, or rather from something else. If these properties stem from "something else" it would suggest that they originate in some underlying mechanism. If, on the other hand, our observations are also often present in these random graphs, it suggests that our findings could be merely artefacts of the data, rather than a reflection of an underlying mechanism.

The first randomisation we consider is based on simply randomising the actual network as we constructed it, which we call the "simple" randomisation. We do this by making a list of all edges with three columns: the first endpoint,

the second endpoint, and the weight of that edge. We then shuffle the column of the first endpoints of the list, so that the links now are randomly connected to other nodes. Since we simply shuffle the list, this will preserve the number of links per node. In addition, the weights remain unchanged, although since the links are no longer attached to the same node, the strength of the nodes are not preserved. In essence, this constitutes a null model to which we compare our empirical results. We generate a hundred samples to obtain our results.

Let us first examine the simple randomisation. Instead of describing the results more elaborately, we briefly summarise the same points as for the empirical network:

- Of course, the degree did not change, so it still varies enormously.
- The average weight per link almost does not depend on the degree.
- The random network is a small world, with an average path length of about 3·34.
- The clustering is much lower and almost does not depend on the degree.
- Hubs do not connect mainly to low degree nodes, and the weight does not play any role.
- The number of common neighbours does not depend on the weight.

Comparing this point by point, we see that the only observation from our empirical network that is also present in this random network is the "small-world" phenomena. This is mainly because random links tend to connect distant parts of the network, which are in some sense "shortcuts". None of the other properties are observed in this random network though, which suggests that our empirical network really is very different from what can be expected in a random network. Of course, this is *with respect to this randomisation*. So, we know that in a random graph with the same degrees and the same weight, we do not expect to see these properties, which suggests that some other underlying mechanism is responsible for these empirical observations.

Of course the empirical network is constructed out of co-occurrences in newspapers, so that is our first "underlying" mechanism. Hence, we create a second type of random graph based on the original co-occurrence data. Similar to the first type of randomisation of the empirical network, we can make a list of links with only two columns: the name of a person, and the sentence in which that name occurs. Similarly, we can shuffle the first column of names. This preserves the number of times somebody occurs in a sentence, and it preserves the number of people that occur in each sentence. After

this randomisation, we apply exactly the same procedure as we did for the empirical graph: We look at the co-occurrence of people in sentences. We call this type of randomisation the "bipartite" randomisation (Newman et al., 2002; Guillaume and Latapy, 2006). Let us summarise the same points as before for this bipartite randomisation.

- The degree and weights are both heterogeneous, which was not part of our constraint in this randomisation.
- The average weight increases with the degree (although slower than empirical).
- The network is a small world, with an average path length of 2.5.
- The (weighted) clustering is equally high, and decreases strongly with degree.
- The neighbour degree strongly decreases with degree, although the weighted neighbour degree does not depend on degree.
- The number of common neighbours clearly increases with the weight.

Hence, the bipartite randomisation shows more observations that are in common with our empirical network. Not all of these measures necessarily correspond quantitatively to our empirical measures, but qualitatively, the bipartite randomisation shows a lot of commonalities with the empirical network.

The conclusion has to be as follows: We must expect to find core-periphery structures in general in networks that are based on co-occurrence. This is so regardless of any actual tendencies of elites to clique together. They differ quite a lot from ordinary random graphs (i.e., simple randomisation), especially concerning the weights. We can, thus, conclude that even though there are some differences, many properties seem to stem from the bipartite nature of the co-occurrence data, rather than from some underlying mechanism. This is a disturbing conclusion. It leads us to suspect that the observed network structure is simply an artefact of the data, rather than a reflection of the structure of the elite in Indonesia. This is not to say that there is no connection between the observed network and social reality. But the connection has to do mainly with the frequency of occurrence (which is sociologically meaningful) and not with the observed pattern of co-occurrence.

Moreover, the clustering of second-rung elites around certain issues or topics is also not reproduced in the bipartite randomisation. Nor does the randomisation reproduce the increase of the weighted neighbour degree with the degree. There are also two large statistical differences: (1) the average degree is much higher in the bipartite randomisation (about 30) than in the empirical network (about 12); and (2) the average weight is much lower (about 1.3) than empir-

ically observed (almost three). Notice that the average strength is about the same in both (about 39 in the bipartite randomisation versus 36 empirically). This is probably a result of the scattering of occurrences over various sentences that ensues the randomisation. There is only a small chance that somebody will randomly occur again in a sentence with the same person. Empirically, this happens much more frequently. People tend to co-occur repeatedly. Consider, for example, Tarmizi Hakim, who co-occurs with only one person, but does so 38 times, rather than once with 38 different people. He was a heart-surgeon who happened to be on board of an aeroplane when a well-known human rights activist, Munir Said Thalib, became ill, and eventually died. This is the only time that Hakim appeared in the news, so that all co-occurrences were only with Munir, never with somebody else.

These differences point to a rather simple model for replicating our network. The exact model is more accurately described in Traag et al. (2014), but we do want to highlight here the key idea. The model is based on two mechanisms: (1) nodes that have many links are likely to attract more links; and (2) if two people already co-occurred, they tend to co-occur again, as suggested by the comparison to the bipartite randomisation. The first mechanism is commonly known as *cumulative advantage* (De Solla Price, 1965), *preferential attachment* (Barabási and Albert, 1999) or the *rich-get-richer* or *Matthews effect* (Merton, 1968). These different names all point to the same idea, those who already have much, gain even more.

It is tempting to associate this observation with society, where the rich also get richer, since that is how power stratifies societies. Once more, however, no such social mechanism is required in this case. Rather, "the rich get richer" here means simply that those who already feature prominently in the news are likely to appear again. In somewhat more detail, the model works as follows. We iteratively add new sentences, and the number of people that will appear in the sentence is randomly chosen, with probabilities similar to the empirical data. We iteratively add persons to this sentence. If we do not add a new person (i.e., somebody that has never yet occurred before), we will add an already existing person. We do so by first looking if another person already occurred in that sentence. If that is not the case, we add a random person, according to cumulative advantage. Otherwise, we pick a random person that already occurred in the sentence, and look at his degree. Then depending on his degree, he will either (A) co-occur with an existing neighbour, or (B) with somebody at random according to cumulative advantage. The balance between these two options A and B is controlled by a single parameter. If option A is more often chosen, people will tend to repeatedly co-occur with the same people, while option B allows to also co-occur with new people.

Notice that this model needs very little empirical input compared to the bipartite randomisation. The bipartite randomisation needs to know how frequently every person appears in the corpus. Our model on the other hand, only needs a couple of things: the number of sentences, how many people appear in sentences, and the balancing parameter. We estimated the parameter by seeing which parameter value best reproduced the average degree and weight.

We briefly highlight the main results of the model:

- The degree, weight and strength distributions are very heterogeneous, and the average degree and weight are almost the same as empirically observed.
- The average weight increases in almost the same manner as empirical.
- The network is a small world with an average path length of 3.36.
- The (weighted) clustering is relatively high, and decreases with degree (although the dependence on degree is somewhat different).
- The neighbour degree increases slightly at first, but then decreases. The weighted neighbour degree increases, however, similar to the empirical.
- The number of common neighbours clearly increases with the weight.

The model thus reproduces quite some of the observations of our network, even though they are not an exact quantitative match. Nonetheless, it is quite impressive that with only such a simple mechanism we can already reproduce some of our observations almost as well, and at times even better, as the bipartite randomisation, which needs much more information. In particular, some of the observations that are not reproduced by the bipartite randomisation, are much more in line with this simple model. In particular, the (weighted) neighbour degree is more closely matched by the model than by the bipartite randomisation. This suggests that these differences between the empirical network and the bipartite randomisation may stem from the fact that people repeatedly occur with the same people. In summary, this simple mechanism can already explain quite a lot of the structure of co-occurrence networks, and should form the basis of future, more realistic models. One of the things that is not taken into account is that most people will appear in the news because of some issue, we presume that taking into account this structure will make the model much more realistic.

In conclusion, although we observe a core-periphery structure in the empirical network, this mainly stems from some simple bipartite mechanism related to the co-occurrence. By comparing our empirical observation to what we can expect based on a bipartite randomisation, we should be able to infer more accurately what observations are significant (i.e., differ from the bipar-

tite randomisation) and what are not (i.e., are almost similar to the bipartite randomisation). In our case, it suggest that few patterns actually stem from the actual elite structure in Indonesia, and rather originates as a by-product of the structure of the data. We thus recommend to carefully take this into account and be aware of what cannot be inferred from such co-occurrence networks.

Conclusion

In this paper, we analysed a co-occurrence network based on a corpus of newspaper articles concerning Indonesia. We have analysed various properties of this network, and at first glance we can characterise it as follows. First, there is a large heterogeneity in the degree of nodes: there are a few major hubs, but the majority are only low degree nodes. Secondly, high degree nodes attract disproportionately much weight, so that the hubs co-occur much more often than their degree justifies. Third, most of the weight is in between these hubs. Fourth, a link is stronger if it shares many common neighbours, consistent with the weak ties hypothesis. Fifth, there is a strong tendency for clustering, with a bias towards stronger links, where hubs show less clustering overall, but the clustering with other hubs is quite strong.

All these characteristics point to a certain core-periphery structure. The core consists of high degree, strong hubs, that are connected mainly by quite strong links. The periphery consists of many disparate local clusters, which are mainly connected to the core. Such a structure resembles the networks we would expect to find among a nation's elites, but the resemblance in this case is misleading. For such a structure appears also in a simple bipartite randomisation, which shows qualitatively largely similar behaviour. That is, the core-periphery structure largely emanates from the co-occurrence structure of the data. To come back to some of our initial questions: Well connected people do not connect significantly more to each other, nor do they connect preferably to less well connected people. Overall, the structure is largely what can be expected from such co-occurrence networks. Some differences with the bipartition randomisation remain though. The clustering around issues or foreign politics is not reproduced by the randomisation. The weighted neighbour degree increases with the degree, different from the randomisation. Also, empirically people tend to occur repeatedly with the same people, which is not the case in the bipartite randomisation.

This last observation points to a simple model we have developed, which also reproduced a core-periphery structure. The model is based on two key

mechanisms: (1) people who occur are more likely to occur again, and (2) people tend to repeatedly occur with the same people. Given its simplicity, the model is very well able to reproduce many of the observed phenomena. In particular, the (weighted) neighbour degree is much closer to what is observed empirically than the bipartite randomisation. This points out that these differences might only stem from the fact that people repeatedly occur with the same people.

Since this type of research can be easily replicated on other corpora, it would be interesting to compare these results with newspapers from other countries and/or times, which may give other results. That latter element is also one of the ignored features in this paper: the temporal dimension. Earlier studies show that links tend to be activated in a bursty manner (Barabási et al., 2005; Malmgren et al., 2008), and that there are intriguing connections between the activation rate and topological properties (Miritello et al., 2011; Karsai et al., 2011; Karsai et al., 2014; Delvenne et al., 2013). Perhaps we can learn something from the temporal analysis of this network.

Some of these temporal variations will likely not be endogenous, but reflect exogenous shocks and developments. For example, in media reports, or Twitter streams, external events have a major impact (Leskovec et al., 2009; Crane and Sornette, 2008). Some extreme events, such as bombings and earthquakes, also show effects on mobile communications networks (Gao et al., 2014; Bagrow et al., 2011). The question is to what extent exogenous events are reflected in the media network. For example, a terrorist attack will suddenly bring certain people in the spotlight, and a rift in a political party might change the structure of co-occurrence. The answers are not yet clear, but these questions are intriguing and merit further analysis.

In summary then, our results suggest that the co-occurrence network derived from this corpus yields relatively little insight into how the elite in Indonesia are connected. It is rather similar to a co-occurrence network if people appear randomly in sentences. This is disappointing, given that we might have hoped to learn whether the elite is centralised or not, or whether they are divided or rather unified. It seems difficult to infer such characteristics based on this type of data. Nonetheless, such a network may provide a helpful tool for exploratory research. Finding out who occurs with whom might provide a quick overview of which people would be interesting for further investigation.

This raises the final question of the future of the technique for mapping elite networks that we have developed in this paper. Is this line of research destined for a niche, but basically not up to the main job? The answer is probably that simple co-occurrence networks of the kind we have investigated are never likely

to yield a sociologically meaningful network structure. This is a significant negative conclusion.

However, if we had more meaningful data about the nature of the relationships between individual elites, the resulting networks may be far more informative. Especially if we can differentiate between different types of social relations it would greatly enrich the methodology. For example, Wasserman and Faust (1994: 37) identify seven basic types of social relationships, which could perhaps be extracted from the texts automatically in the future. Our present work seeks to extract such information from the newspaper data, still using digital methods.

References

Amaral, L.A.N., A. Scala, M. Barthélémy and H.E. Stanley (2000) "Classes of small-world networks". *Proc. Natl. Acad. Sci. USA* 97: 11149–11152.

Bagrow, J.P., D. Wang and A.-L. Barabási (2011) "Collective response of human populations to large-scale emergencies". *PLoS ONE* 6: e17680.

Barabási, A.-L. (2009) "Scale-Free Networks: A Decade and Beyond". *Science* 325: 412–413.

Barabási, A.-L. and R. Albert (1999) "Emergence of Scaling in Random Networks". *Science* 286: 509–512.

Barabási, A.-L., A. Bees and N. York (2005) "The origin of bursts and heavy tails in human dynamics". *Nature* 435: 207–211.

Bresis, E.S. and P. Temin (2008) "Elites and Economic Outcomes", in in Steven N. Durlauf and Lawrence E. Blume (eds.) *The New Palgrave Dictionary of Economics*. 2nd edition. Basingstoke: Nature Publishing Group.

Bullmore, E. and O. Sporns (2009) "Complex brain networks: Graph theoretical analysis of structural and functional systems". *Nat. Rev. Neurosci.* 10: 186–198.

Burton, M.G. (1984) "Elites and Collective Protest". *Sociol. Q.* 25: 45–66.

Crane, R. and D. Sornette (2008) "Robust dynamic classes revealed by measuring the response function of a social system". *Proc. Natl. Acad. Sci. USA* 105: 15649–15653.

Cranmer, S.J., E.J. Menninga and P.J. Mucha (2014) "Kantian fractionalisation predicts the conflict propensity of the international system". *arXiv*:1402.0126 [physics].

Csermely, P., A. London, L.-Y. Wu and B. Ussi (2013) "Structure and dynamics of coreperiphery networks". *J. Compl. Net.* 1: 93–123.

De Solla Price, D.J. (1965) "Networks of Scientific Papers". *Science* 149: 510–515.

Delvenne, J.-C., R. Lambiotte and L.E.C. Rocha (2013) "Bottlenecks, burstiness, and fat tails regulate mixing times of non-Poissonian random walks". *arXiv*:1309.4155 [cond-mat, physics: physics].

Dogan, M. and J. Higley (1998) *Elites, Crises, and the Origins of Regimes*. Lanham: Rowman & Littlefield Publishers.

Eff, E.A. and P.W. Routon (2012) "Farming and Fighting". *Structure and Dynamics* 5.

Finkel, J.R., T. Grenager and C. Manning (2005) "Incorporating non-local information into information extraction systems by gibbs sampling". Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA. Pp. 363–370.

Gao, L., C. Song, S. Gao, A.-L. Barabási, J.P. Bagrow and D. Wang (2014) "Quantifying Information Flow During Emergencies". *Sci. Rep*. 4, 10.1038/srep03997.

Garlaschelli, D., G. Caldarelli and L. Pietronero (2003) "Universal scaling relations in food webs". *Nature* 423: 165–168.

Garlaschelli, D. and M.I. Loffredo (2005) "Structure and evolution of the world trade network". *Physica A* 355: 138–144.

Goldstone, J.A. (1991) *Revolution and Rebellion in the Early Modern World*. University of California Press.

——— (2001) "Toward a fourth generation of revolutionary theory". *Annu. Rev. Polit. Sci*. 4: 139–187.

Gonsáles, M.C., C.A. Hidalgo and A.-L. Barabási (2008) "Understanding individual human mobility patterns". *Nature* 453: 779–782.

Granovetter, M. (1973) "The Strength of Weak Ties". *Am. J. Sociol*. 78: 1360–1380.

Guillaume, J.-L. and M. Latapy (2006) "Bipartite graphs as models of complex networks". *Physica A* 371: 795–813.

Guimerà, R., M. Sales-Pardo and L.A.N. Amaral (2007) "Classes of complex networks defined by role-to-role connectivity profiles". *Nat. Phys*. 3: 63–69.

Guimerà, R., D.B. Stouffer, M. Sales-Pardo, E.A. Leicht, M.E.J. Newman and L.A.N. Amaral (2010) "Origin of compartmentalisation in food webs". *Ecology* 91: 2941–2951.

Hagmann, P., L. Cammoun, X. Gigandet, R. Meuli, C.J. Honey, V.J. Wedeen and O. Sporns (2008) "Mapping the structural core of human cerebral cortex". *PLoS Biology* 6: e159.

Hicks, J., Reinanda R., Traag, V.A. (2015, in press) "Old Questions, New Techniques: A Research Note on Identifying Political Elites Computationally". *Comp. Sociol*.

Higley, J. and M. Burton (2006) *Elite Foundations of Liberal Democracy*. Lanham: Rowman & Littlefield Publishers.

Joshi, D. and D. Gatica-Peres (2006) "Discovering Groups of People in Google News". Proceedings of the 1st ACM International Workshop on Human-centered Multimedia, HCM '06, ACM, New York, USA. Pp. 55–64.

Karsai, M., M. Kivelä, R. Pan, K. Kaski, J. Kertéss, A.-L. Barabási and J. Saramäki (2011) "Small but slow world: How network topology and burstiness slow down spreading". *Phys. Rev. E* 83, 10.1103/PhysRevE.83.025102.

Karsai, M., N. Perra and A. Vespignani (2014) "Time varying networks and the weakness of strong ties". *Sci. Rep*. 4, 10.1038/srep04001.

Knoke, D. and J.H. Kuklinski (1982) *Network analysis*. Sage Publications.

Leskovec, J., L. Backstrom and J. Kleinberg (2009) "Meme-tracking and the Dynamics of the News Cycle". Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, ACM, New York, USA. Pp. 497–506.

Malmgren, R.D., D.B. Stouffer, A.E. Motter and L.A.N. Amaral (2008) "A Poissonian explanation for heavy tails in e-mail communication". *Proc. Natl. Acad. Sci. USA* 105: 18153–18158.

Maos, S., L.G. Terris, R.D. Kuperman and I. Talmud (2008) "What Is the Enemy of My Enemy? Causes and Consequences of Imbalanced International Relations, 1816–2001". *J. Politic*. 69: 100–115.

Merton, R.K. (1968) "The Matthew effect in science". *Science* 159: 56–63.

Mills, C.W. (2000) *The Power Elite*. Oxford University Press.

Milne, D. and I.H. Witten (2008) "Learning to Link with Wikipedia". Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08, ACM, New York, USA. Pp. 509–518.

Miritello, G., E. Moro and R. Lara (2011) "Dynamical strength of social ties in information spreading". *Phys. Rev. E* 83, 045102.

Newman, M. (2010) *Networks: An Introduction*. Oxford University Press.

Newman, M.E.J., D.J. Watts and S.H. Strogats (2002) "Random graph models of social networks". *Proc. Natl. Acad. Sci. USA* 99 (Suppl. 1): 2566–2572.

Ösgür, A. and H. Bingol (2004) "Social Network of Co-occurrence in News Articles", in C. Aykanat, T. Dayar, and I. Korpeoglu (eds.) *Computer and Information Sciences—ISCIS 2004, Lecture Notes in Computer Science No. 3280*. Heidelberg: Springer Verlag, Pp. 688–695.

Petri, G., M. Scolamiero, I. Donato and F. Vaccarino (2013) "Topological Strata of Weighted Complex Networks". *PLoS ONE* 8, e66506.

Pouliquen, B., H. Tanev and M. Atkinson (2008) "Extracting and learning social networks out of multilingual news", in Proceedings of the social networks and application tools workshop (SocNet-08) pp. 13–16. Skalica, Slovakia, 19–21 September 2008.

Putnam, R.D. (1976) *The Comparative Study of Political Elites*. Prentice Hall.

Reinanda, R., M. Utama, F. Steijlen and M. de Rijke (2013) "Entity network extraction based on association finding and relation extraction", in in Trond Aalberg, Christos Papatheodorou, Milena Dobreva, Giannis Tsakonas, Charles J. Farrugia (eds.) *Research and Advanced Technology for Digital Libraries*. Springer Berlin Heidelberg, Pp. 156–167.

Simini, F., M.C. Gonsáles, A. Maritan and A.-L. Barabási (2012) "A universal model for mobility and migration patterns". *Nature* 484: 96–100.

Steinberger, R. and B. Pouliquen (2007) "Cross-lingual Named Entity Recognition". *Ling. Inv*. 30: 135–162.

Thomas, W.I. and D.S.T. Thomas (1928) *The Child in America: Behavior Problems and Programs*. Johnson Reprint Corporation.

Traag, V.A., R. Reinanda and G. van Klinken (2014) "Structure of an elite co-occurrence network". *arXiv*:1409.1744 [physics].

Traag, V.A., P. van Dooren and Y. Nesterov (2011) "Narrow scope for resolution-limit-free community detection". *Phys. Rev. E* 84, 016114.

Travers, J. and S. Milgram (1969) "An experimental study of the small world problem". *Sociometry* 32: 425–443.

Turchin, P. (2005) "Dynamical Feedbacks between Population Growth and Sociopolitical Instability in Agrarian States". *Structure and Dynamics* 1.

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

Watts, D.J. and S.H. Strogats (1998) "Collective dynamics of 'small-world' networks". *Nature* 393: 440–442.