BRILL

ASIAN JOURNAL OF
SOCIAL SCIENCE 43 (2015) 567–587

Asian Journal
of Social Science
brill.com/ajss

Turning Digitised Newspapers into Networks of Political Elites*

Jacqueline Hicks
KITLV/Royal Netherlands Institute of Southeast Asian and Caribbean Studies (primary author)

Vincent A. Traag
KITLV/Royal Netherlands Institute of Southeast Asian and Caribbean Studies (network analysis)

Ridho Reinanda
ISLA, University of Amsterdam, The Netherlands (information extraction)

Abstract

This paper introduces the Elite Network Shifts (ENS) project to the Asian Studies community where computational techniques are used with digitised newspaper articles to describe changes in relations among Indonesian political elites. Reflecting on how "political elites" and "political relations" are understood by the elites, as well as across the disciplinary boundaries of the social and computational sciences, it suggests ways to operationalise these concepts for digital research. It then presents the results of a field trip where six Indonesian political elites were asked to evaluate the accuracy of their own computational networks generated by the project. The main findings of the paper are: (1) The computational identification of political elites is relatively successful, while much work remains on categorising their relations, (2) social scientists should focus on capturing single dimensions of complex social phenomena when using computational techniques, and (3) computational techniques are not able to capture multiple understandings of social concepts.

* All authors are researchers with the Elite Network Shifts project. The work on which this paper is based has been supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, under the auspices of the Elite Networks Shifts project. For further information, see http://ehumanities.nl. Many thanks to Gerry van Klinken and two anonymous reviewers who provided valuable insights on previous drafts.

 | DOI: 10.1163/15685314-04305004

Keywords

Indonesia – computational methods – political elite – political relations – networks

Introduction

Computational research is in some ways anathema to area studies specialists. Our discipline places value on the type of detailed knowledge that can only be learned from a period of in-country immersion. By contrast, computational research can seem far removed from the social reality it seeks to describe—creating bite-sized chunks of data, cleansed of the messy contradictions of everyday life for processing by computers.

But whether it is through the use of Google, or newly coded tools for specific corpora, such as presented in this paper, we all already use the computational filtering of digital data to make sense of the world around us. Understanding how the techniques work and critically, but constructively, engaging with them is the job of the social sciences.

This paper does this in three main ways. The first section introduces the Elite Network Shifts project to the Asian Studies community, describing the corpus of digitised newspapers and the methodological steps undertaken. The purpose of this section is to lay bare the nature of the data used and to clarify exactly what the computational techniques involve in non-technical terms. Computational methods are sometimes criticised for obscuring their workings into a "black box", unknowable to the uninitiated. This section tries to open that black box.

The following section then discusses how the social science concepts of "political elite" and "political relations" could be operationalised for computational research. Analysing large bodies of text with computational techniques often produces results that are difficult to relate to mainstream social science questions. Rather than working inductively from the bottom up—trying to build significance for bigger questions from the granular wordcount level, this paper works deductively from the conceptual level down. The digital turn *does* give the social sciences new vigour, but not only because it provides new types of beautifully visualised empirical nuggets to help build arguments. The renewed vigour is also because computational techniques *necessitate* qualitative social scientists to carefully consider, and explain, the assumptions behind categories we sometimes take for granted.

The final section presents the results of a fieldwork trip to Indonesia aimed at verifying the validity of the elite networks found using the computational

techniques. It not only introduces a novel way to evaluate the findings of computational research, but also highlights the kinds of difficulties involved in making truth claims about social phenomena when the objects of study are themselves self-reflecting humans (Flyvbjerg, 2002). The paper then concludes with some reflections on how to best approach the use of computational techniques to answer the kinds of questions that interest social scientists.

The Elite Network Shifts Project

Overall Goals

The Elite Network Shifts project (ENS)[1] was conceived to exploit the increasing availability of digitised newspaper articles and advances in computer techniques to automatically extract and visualise information. Initially, the project's goals were to uncover political elite behaviour at times of regime change. It aimed to speak to the literature on elite circulation, comparing what social scientists have already said about Indonesia's two major regime transitions of 1945 and 1998 to the information found using the new computer techniques. However, the experimental nature of the project also made room for investigations into other ways that the sources and techniques could be used to answer questions about Indonesian political development more generally. We began by seeking how far we could locate a set of political elites and describe their relations, then see where this information can be used in social science debates about political elite relations.

Usually, digital data in the social sciences is used to describe the electronic traces left by our communications on social media, email, mobile phone records and geographic location. The kinds of social science questions such data is used for range from analyses of how riots spread using twitter posts (Procter et al., 2013), to studies of human mobility from mobile phone data (Phithakkitnukoon et al., 2012) and public involvement in petition signing (Margetts et al., 2013).

By contrast, the ENS project does not use the data contained in the hundreds of thousands of newspaper articles to represent something about the person who created the text. Most projects involved in the computational analysis of relatively long digitised texts, such as media articles or books, are interested

1 The project, which began in September 2012 and will run until 2016, is funded by the Royal Netherlands Academy of Arts and Sciences (KNAW) through the Computational Humanities Programme. It includes researchers from KITLV, NIOD, University of Amsterdam and Erasmus University. For more information, see http://kitlv.nl/research-projects-elite-network-shifts/.

in some sort of textual discourse, describing textual features to say something about the writing itself. Grouped under the general heading of "text mining", the computational methods can do several different types of analysis. They can trace the historical development of words and phrases, like Google's *n-gram viewer*[2] and track the relationships of several different words over time to help map conceptual development. They can distinguish texts by author or genre based on writing style by clustering different linguistic features. They can group related words that appear together in the same documents to produce the "topics" discussed in a text. In the social sciences, the use of such techniques has included the analysis of presidential debates (Martin, 2012), proposed legislation hearings (Adler and Wilkerson, 2011) and political treaties (Spirling, 2012).

ENS tried to go one step further than these discourse analyses. Based on the proposition that it is common practice for social scientists to use newspaper articles extensively in their qualitative research, we wanted to explore whether the computational analysis of newspapers could also be used to describe social realities. We did so critically, understanding that the inferential leap from reading something in a newspaper to a real occurrence is a big one.

Newspaper accounts differ from the reality they seek to describe in so many fundamental ways that a whole field of enquiry has grown up around them, with studies explaining their discourses in terms of political bias, professional ideology and commercial context among others. Indeed, one part of this three-year project interrogated the media itself to look at the framing of stories from different newspapers, their timing and topic selection. Nevertheless, accepting all the limitations of newspapers as a source that researchers with non-computational methods do, we searched for what newspaper articles could tell us about relationships among the Indonesian political elite.

Digitised Corpora

We collected several different corpora of newspaper articles during the course of the project, but only one was used for initial experimentation: the *Joyo* corpus. This was compiled by a news service called *Joyo*, started by Gordon Bishop in 1996, whose original aim was to circulate information on Indonesian politics at a time of press restrictions under an authoritarian regime.[3] It included over 140,000 mostly English language articles on Indonesian politics from 1998–2012. The *Jakarta Post* and international media predominate, but there were also some translated articles from Indonesian newspapers.

2 https://books.google.com/ngrams.

3 http://www.joyonews.org.

In our "data bank", there were several additional corpora. The first was 220,000 articles from another compilation of articles called *Islands* which had been collected by the ENS project's co-ordinator, Gerry van Klinken, using web-crawling software on a chosen set of news websites. It contained a mixture of Indonesian and English language articles ranging from 1990–2009. *Apakabar* is a second additional corpus of around 175,000 articles collected by John A. MacDougall and maintained by Ohio University.[4] It includes Indonesian and English language articles from 1990–2002.

A third collection of 530,000 articles was bought from a media-scanning company in Indonesia. All in Indonesian and covering 2011–2013, it was the most complete corpus we had covering most stories from 25 regional and 25 national newspapers. For historical newspapers, the National Library of the Netherlands (KB) has digitised thousands of editions of Dutch colonial newspapers going back to 1618[5] and the Institute for War, Holocaust and Genocide Studies (NIOD) holds a comparatively smaller collection of Indonesian newspapers from the period 1930–1950.[6]

In Indonesia, the National Library of Indonesia (PNRI) has yet to digitise its large collection of newspapers. A few media organisations have full sets of their own newspapers digitised from around 2010 onwards, with the exceptions of *Kompas* and *Majalah Tempo*, which have digital archives stretching back to 1965 and 1974 respectively. While articles from the latter two can be downloaded individually for a fee, neither company was willing to sell their full database in a form conducive for computational analysis for the project's internal use. Some commercial news services, such as Lexis Nexis, carry a few Indonesian newspapers from around 2010 onwards, but limit their downloads.

A final alternative for anyone interested in the availability of digitised articles on Indonesia is to begin the automatic collection of *current* news themselves using a web scraping programme. While comprehensive, this does not tend to ensure *complete* coverage of a news organisation's output, as it often misses some articles and collects only the online summaries of others.

Methodological Steps

Projects using computational techniques either employ social scientists who can code or put together multi-disciplinary teams. While the latter has the benefit of a higher level of computational expertise, it becomes more important

4 http://www.library.ohiou.edu/indopubs/.

5 http://kranten.delpher.nl/.

6 http://niod.x-cago.com/maleise_kranten/papers.do.

for the non-coding social scientists to understand the many technical decisions that go into the development of computational tools.

As Gooding notes, "When texts are deconstructed to the extreme of granularity, and interactions become mediated by automated tools [...] there is merely a massive corpus of words which carry no great epistemological significance" (2012: 4). Computational results nearly always *seem* to say something because they are made up of combinations of words which the researcher recognises, so a basic understanding of the methodological steps is needed to effectively check their validity. It is also useful to realise how much manual work is needed to set up these automated processes. The following is a basic categorisation of the types of tasks undertaken for computational analysis of unstructured text (that is, text like articles, books or even tweets where the text is not yet sorted according to certain features into structured tables).

1. *Data collection*: Digitised newspaper articles should be born-digital or scanned using sufficient quality Optical Character Recognition (OCR) technologies. Ideally, each article is contained in a single file rather than gathered together in one file per newspaper edition since stories often run across different pages and their integrity can be otherwise lost.
2. *Data pre-processing*: The amount of pre-processing depends on how well-structured the text is. The main idea behind the pre-processing stage is to turn unstructured raw data into a shorter, structured format that maintains all the characteristic features of the original text without losing accuracy. This step includes:
    a. *Indexing*: Words are sorted into indices, just like in a book, to enable the programme to search faster. Depending on the task the index will be used for, some decisions can be made to exclude some words, such as "the" or "and".
    b. *Tagging*: Each word or punctuation mark is automatically labelled as a part of speech, such as noun, verb and adjective.
    c. *Parsing*: Can be thought of as a tree where each of the connecting leaves is made up of several of the tagged words together. Labelled adjectives ("brown") and nouns ("fox") become noun phrases ("brown fox"). There is a wide range of ways to parse, running from "shallow" to "deep", the latter implying much more manual work than the former.
3. *Information extraction*:
    a. *Named entity recognition*: This technique automatically detects and classifies names, places and organisations based on signals in the text like capitalisation and sentence position. It is one of the more successful techniques, generating relatively accurate results.

b. *Disambiguation*: Often a single person will be referred to in the text by several different names. For example, "Susilo Bambang Yudhoyono" can also be referred to as "President Yudhoyono" or "SBY" or may be misspelled as Yudoyono. The process of disambiguation ensures that all these different ways of referring to one person are grouped together, effectively recognising them as the same person. In order to do this, we used "string similarity" techniques which recognises statistically significant similarities in the order of letters between two different spellings of a name along with "Wikification", which automatically uses the lists of different versions of a person's name in their entries on Wikipedia.

c. *Filtering*: Since we were, at that first stage, only interested in relations among elites, we set aside the organisations and place names for later use, and focused on people's names. Because we defined elites as people of influence, we assumed that a person would be influential if they appeared in the articles more often than average. We further discarded all names that did not appear more than the average in our corpus.

4. *Network creation*:

a. *Co-occurrence*: A "computational relation" was created when a frequently occurring person appeared in the same sentence together with other frequently appearing people. These were people with a high degree of "network centrality".

b. *Clustering*: The process of grouping persons who have more connections to each other than the other persons. Decisions are made here about the type of clustering algorithm used, as well as their parameters.[7] We used this technique to isolate groups of Indonesian political elites from other types of people appearing in the media.

c. *Visualisation*: These computational relations were visualised using igraph—an open source network analysis package. Figure 1 is an example of a visualisation, showing a clustering of different groups of elites based on how frequently they co-occur together.

Operationalising the Concepts of "Political Elites" and "Relations" for Computational Research

The computationally-derived networks are thus constructed with names extracted from the articles, and based on frequency of appearance, co-occurrence

7 Here, Traag used the CPM method, in contrast to most who use modularity.

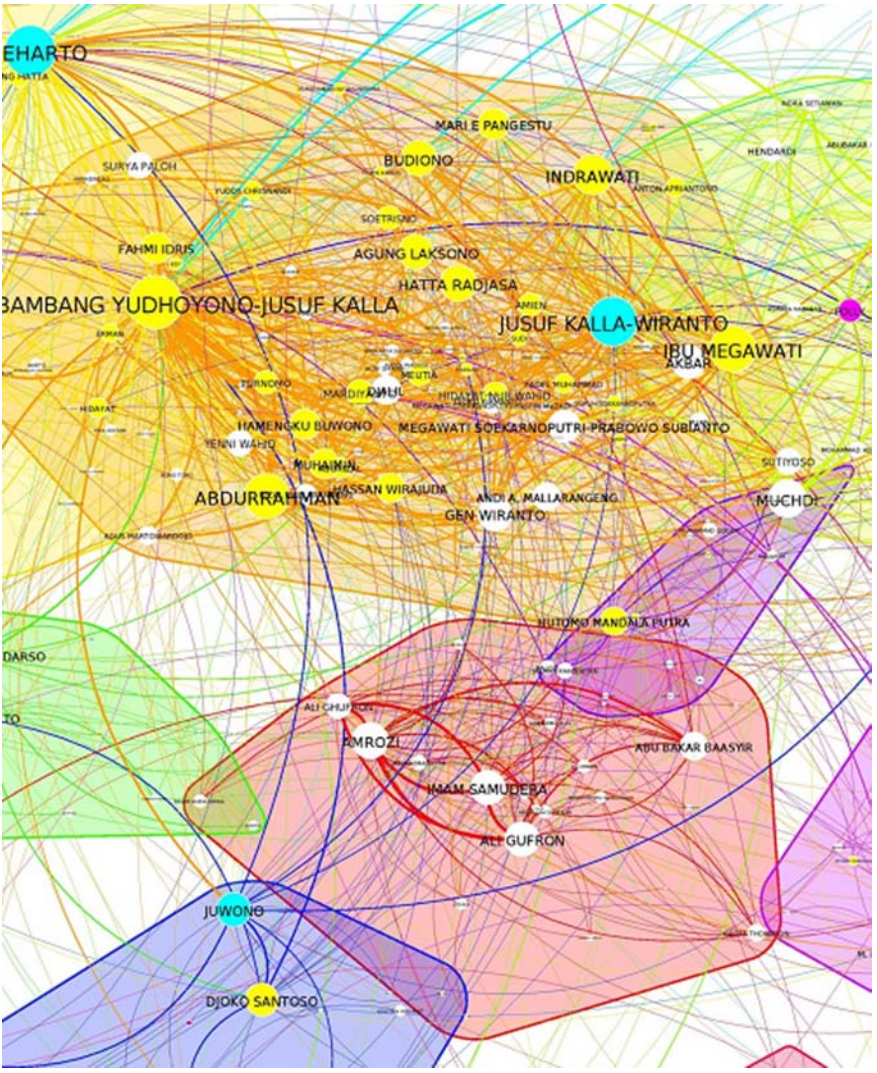

FIGURE 1 *A network visualisation of clusters of Indonesian political elites*

together within the space of one sentence, and network centrality. Even though all of these techniques rely on sophisticated underlying mathematics, the results are a crude representation of all the complex and nuanced elements that exist in a real relationship between political elites. However, by applying some of the insights from the literature on political elite relations, we begin to better understand the methods' potential and limitations in social science terms.

When looking at relations among political elites, we must first find ways to identify populations of elites before describing their relations. Following

Putnam's (1976) analysis, any researcher attempting to find a population of political elites has a choice between three main methods. These are:

1. *Positional*: Defines political elites as those in "command positions" at the top of major political institutions (Scott, 2003). The researcher defines for herself which institutions to include, as well as the level of seniority within each institution as these change depending on the different country and historical contexts.
2. *Decisional*: Defines elites as those whose "preferences regularly prevail in cases of differences in preferences on key political issues" (Dahl, 1958: 464). Rather than identifying elites from the position lists of various institutions, the decisional method uses documents and interviews related to a particular decision-making process. It, therefore, depends on the researcher's choice of policy issues, as well as the judgements of either the researcher or insiders about whose decision prevailed.
3. *Reputational*: Uses elements of both the positional and decisional, by first identifying a number of elites based either on their positions or the advice of an expert panel, then expanding this selection by asking those elites to nominate others.

Broadly, the decisional and the reputational approaches are better able to uncover informal and overt relations of power since they study actual decision-making processes. By contrast, the positional method accesses formal relations and the *potential* for power since it assumes that power is tied to the resources associated with position and institution.

These three methods are the building blocks of the more conceptual definitions of political elites that are often found in the literature, such as Higley's (2009: 163) commonly used one: "Elites may be defined as persons who, by virtue of their strategic locations in large or otherwise pivotal organisations and movements, are able to affect political outcomes regularly and substantially." Here, Higley's definition is clearly positional, and by referencing both organisations and movements, expands the normal focus on government and legislatures into the realm of civil society. Another part of his definition, the "ability to affect political outcomes" also includes something of the decisional method's insistence on *successful* outcomes of the elites' exercise of power.

With all of these issues in mind, locating political elite with our computational method started by distinguishing political elites from other types of frequently mentioned persons in the media (e.g., sports or entertainment personalities) using clustering, as shown in Figure 1. It was based on the principle that, for example, sports celebrities will usually be mentioned together in sen-

tences with other sports celebrities, rather than with politicians. This allowed us to quickly identify and manually discard any groups that were not national political elites.

A more detailed comparison of this computational method's results with a manually-built list of positional elites is described elsewhere (Hicks et al., forthcoming). Suffice to say here that it both reflects the more established three methods and adds something new. In terms of similarities, the clusters of elites uncovered by it tend to occur around specific issues that come close to the policy domains of the decisional method. In choosing which communities to include as national elites after the community detection analyses have been performed, we draw on the type of knowledge of institutions that is required by the positional approach. I (the primary author) would be alarmed, for example, if cabinet members did not show up in our elite population since I consider these positions to be ones of political influence. Elite reputations are reflected in our method to the extent that persons will be interviewed by journalists based on their reputation as experts on the issue being discussed or will appear as subjects of stories partly because of their importance.

The differences between the computational method and the previous three more established ones can be seen in its underlying assumptions. Our method requires no decisions from the researcher about which institutions and positions to include, nor which policy domains. The boundaries of an elite population are not presumed, but are set by decisions about the frequency and proximity of co-occurrence in newspaper articles, with all of the difficulties which that implies. The nature of newspaper coverage is also that it tends to cover relatively sensationalistic stories. This meant that our use of *frequent* co-occurrences uncovered only those types of "relations" involved in sensationalistic stories like corruption scandals or legal disputes, missing the sort of behind-the scenes relations that play such a crucial role in Indonesian politics. It also became apparent during the course of the project's first year that it was the underlying bi-partite structure of co-occurrence in one sentence that gave the overall network structure its shape, rather than anything connected to the actual information contained in the sentences (see Traag's paper in this edition for the technical details).

Returning to Higley's definition of a political elite as those who are able to "affect political outcomes regularly and substantially" (2009: 163), the regularity is captured by frequency of co-occurrence and the "substantially" by the idea that newspapers provide information about people or events of public interest. In short, the computational method has proven a relatively practical way of implementing some of the elements of Higley's definition.

Having found a working definition of political elites, the next step was to define the types of elite relations that interest social scientists and match them to the kinds of indicators that can be found computationally in newspaper articles. A review of a portion of the academic literature on political elite relations reveals that the kinds of relationships that interest social scientists can be broadly interpreted in three main ways: (1) political endorsement or opposition, (2) patron-client relations, and (3) ideological. To what extent can these types of relations be uncovered using the computational techniques?

Political endorsement or opposition can be placed in a wider literature that describes the level of "politicking" among a small group of political elites. Stories of dismissals, promotions, and various battles among shifting groups of elites for control over powerful institutions and their resources are at the centre of these types of analyses. Where endorsement or opposition of one elite for another is for public office, whether it is to head a particular institution or gain a seat in parliament, such relations are very likely to be reported on extensively in the media, therefore holding relatively strong promise for those working on digitised news corpora.

The second type of relationship is patron-client. At its heart is an exchange of an economic good for political support between someone with authority or wealth (patron) and another who benefits from them (client). In developed countries, such exchanges are often studied as the "soft corruption" of corporate lobbying or political donations, while for developing countries the emphasis is more likely to be on "hard corruption" and/or the social and cultural significance of clientelist systems. There are some creative ways such relations could be approached using computational techniques, such as exploring correlations of individual political donations and sentiment analysis[8] of the beneficiary politician or party towards the source of the donation. Nevertheless, and particularly in a country where there is a high incidence of informal political donations, their clandestine nature make these types of relations ill-suited to computational information extraction from news articles.

The third type of relation—ideological position—is somewhat amenable to the computational techniques of topic modelling. Here, common co-occurrences of groups of words are found so that topics of individual documents can be inferred.[9] Then, connection of these topics to elite names and the application of sentiment analysis techniques could give elite positions on issues.

8 Sentiment analysis techniques try to determine opinions and attitudes based on polarising words, such as "support" or "disagree".

9 Scott Weingart explains topic modelling very well (see http://www.scottbot.net/HIAL/?p=19113).

However, both topic modelling and sentiment analysis for these more nuanced opinions do not yet work well (Balahur and Steinberger, 2009). In principle, as we ask for more complexity from the computational methods, their accuracy correspondingly falls.

At this point in our project, the types of relations that co-occurrence alone represented were simply too diverse to ascribe any general meaning. For example, they can be two people who comment on the same issue, colleagues in the same organisation, competitors for the same post, have attended the same meeting, are locked in a court battle, or are members of the same family.

To try to fill the gap between what social scientists find interesting and what the computational techniques can do, my computational colleagues are currently experimenting with extracting some sort of meaning from the co-occurrence by interrogating its linguistic context. At the time of writing, they were initially looking for indicators of (1) functional relationships, such as A works for B or C is the daughter of D, and (2) topic relations, such as A and B are proponents of anti-corruption legislation (Reinanda, 2014).

From the social science side, it may be useful to understand computational tools as fundamentally an exercise in classification. This makes the job of the social science researcher one of framing questions that can use the kind of results that computational techniques are capable of delivering. Using the example of a relatively complex, interpretive relation, such as elite unity or disunity over a period of regime transition, we can see how this question may be broken down to elements that the computational methods are more likely to find.

Most studies of elite unity over an authoritarian to democratic regime transition describe unity in terms of support or opposition for the leader of the government and/or other elites who are challenging that position. But that unity can itself be described in any number of ways: from the level of consensus over institutional forms (Burton and Higley 1987); the level of competition encouraged among candidates fielded by political parties for elections (Zavala, 2013); ideologies about the division of authority within a political system (O'Donnel and Schmitter, 1986); and prioritisation of reform issues, such as corruption or rules that allow inter-party competition (Abente, 2009). All of these can be approached in one way or another with computational tools resulting in a level of accuracy that reflects not only the development of the techniques in the wider computer science community, but also the level of manual input that each researcher is prepared to contribute.

Topic modelling coupled with sentiment analysis are theoretically suited to uncovering fragmentation or cohesion of elite relations in whatever terms the researcher describes them. However, particularly for newspaper content, their

success relies on even more tightly focused questions than those above. For this example, a relatively simple sentiment such as "endorsement" or "opposition" is beginning to prove workable, as has already been shown in some studies (Sudhahar et al., 2012). This shows how important it is to be realistic about what computational techniques can do. Even under the best circumstance, they give results that may be able to computationally augment social science research, but not replace other methods or types of analysis.

Field Trip to Check Validity of Results in Indonesia

While my ENS colleagues were working on finding ways to extract what sort of relationships co-occurrence indicated, I (the primary author) undertook a field trip to Indonesia to compare our initial computational networks to the elites' own perceptions of their network.

Usually, the accuracy of computational results are evaluated in relation to the text. At this stage of the ENS project, our goal was to automatically identify names which co-occur within one sentence and, referring back to the text, we found that the computational tools were able to do this with a high degree of accuracy.

Otherwise, projects often work with "domain specialists"—people from the social sciences or humanities who decide whether the computational results seem to make sense. Since this project looked at fairly contemporary relations, rather than historical ones,[10] we had the luxury of being able to ask the subjects of the analysis themselves how far our results matched their own perceptions of their networks.

We could have checked the results against analyses of elite relations from the secondary literature on Indonesian politics, but coverage of all the elite relations that showed up in each network is not complete. It also seemed more compelling to go from the extremes of such an abstracted stylisation of elite relations to the people themselves. This, coupled with a basic curiosity about how the elites would react to seeing their computationally-derived networks and the degree to which they would agree their accuracy, took me to Indonesia.

Such an exercise was fraught with methodological complications since it tested not just the accuracy of the computational techniques, but also the

10 For example, see Text To Political Positions (http://www2.let.vu.nl/oz/cltl/t2pp/); Biography Net (http://www.biographynet.nl/biographical-data-in-a-digital-world/).

reliability of newspaper accounts, as well as elite self-reflexivity and honesty. In practice, it is impossible to systematically disentangle their effects.

For this part of the project, we took a positivist approach to the understanding of newspaper stories, assuming they were able to transmit a set of *facts* about the world. This was in contrast to later research questions that took a more constructivist approach to understanding and exploring news stories as *narrative*.

The reliability of the elites' own accounts of their networks also plays a mediating role in any findings from this field trip exercise. It is not possible to fully accept the results of the elite interviews as a kind of "gold standard" of actual elite relations. As we will shortly see, there is a strong possibility that the elites answered questions about their own relations based on how they perceive themselves, or how they wish to be perceived by others. This could have been explored to an extent by showing the elites two networks—one randomised and one based on our co-occurrence technique.[11] It is common practice in statistical correlations to compare results to a randomised output and a useful lesson to learn when engaging in quantitative research.

Table 1 below introduces the interviewees. I tried to choose people who appeared relatively frequently in articles and had a good number of links—at least 25. Otherwise, in addition to aiming for people from different sectors, the final list was based on purely pragmatic concerns of who I was able to contact and agreed to meet me. Other area specialists on Indonesia broadly agreed that these interviewees could be considered political elites.

The computational networks I showed to the interviewees were derived from the *Joyo* corpus only since that was the first and easiest to pre-process as it was mostly in English. Rather than showing the interviewees a complicated network including multiple levels of links, such as is shown in Figure 1, I showed them only a visualisation of their direct links, known as a "star" or "ego" network. If there were more than around 30 direct links, I reduced them, discarding the least frequent links so as not to burden the interviewees with very long lists. Each network covered a specific time frame since the project is interested in change over time. This was mostly the last three years of data (2010–2013), but in one interview I extended it to ten years to generate a higher number of potential links.

An example of the network I showed to one interviewee is shown below in Figure 2. This particular one covered the years 2010–2013.

11 Thanks to Tom Pepinsky for raising this at the AAS presentation.

TABLE 1 *Fieldwork interviewees*

Name	Category	Description
Akbar Tandjung	Politician	An influential politician in the Golkar Party since the mid-1980s.
Amien Rais	Religious-political leader	Former head of Muhammadiyah; the political face of the early stages of the *reformasi* movement.
Emerson Yuntho	Civil-society activist	One of the heads of a well-known civil society advocacy organisation, Indonesian Corruption Watch.
Agus Widjojo	Ex-military general	Former chief of the TNI Territorial Affairs and a leading reformer within the military.
Sofyan Wanandi	Businessman	Owner of Gemala Group; long-time spokesperson for Indonesian business interests.
Faisal Basri	Election candidate	University economist; candidate for Jakarta elections in 2007.

Although I initially showed them the visualisation of the network as above, we ended up working through lists of names. It is possible that other types of visualisations, such as word clouds, could have had some effect on their answers, but in the limited time given for interviews, this was not something I decided to test.

I asked them three main questions about each person that appeared in their computational network.

– Have you frequently communicated with this person in the last three years?
– Is this person in your political network, as you define it yourself?
– How would you describe your relationship with this person?

The first question about whether or not they had communicated with each person in their computational network came from the literature on social network analysis. Most studies using social network software to visualise human relations are based on survey questions, rather than computational extraction, and a question about communication is the usual way that a relationship is

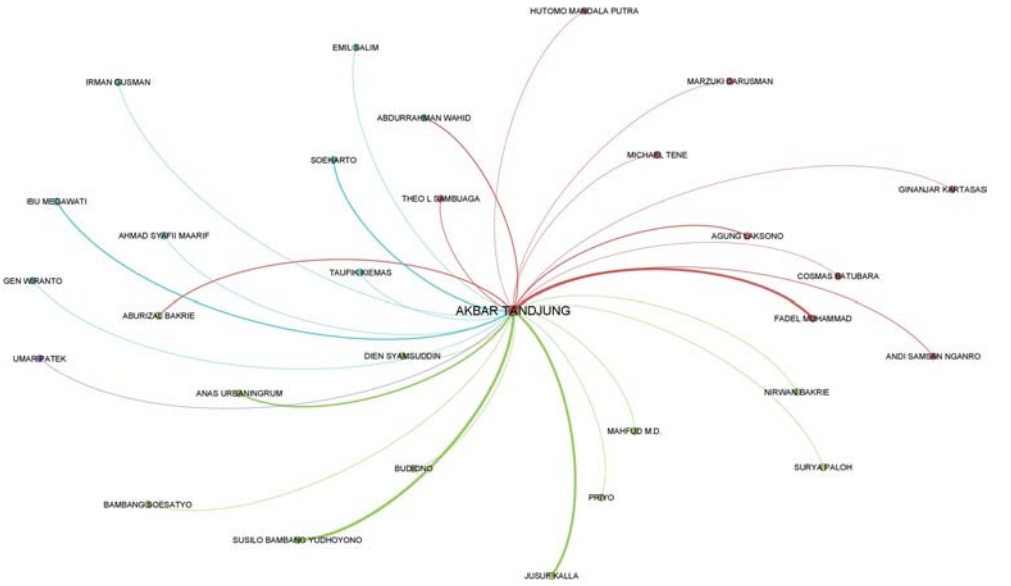

FIGURE 2 *An example of a computational network shown to interviewees*

defined in such surveys. The overlap between those persons an elite had communicated with and those that they confirmed as being in their network as defined by themselves was very high—92%.

Asking the interviewees to use their own definitions of their political network were aimed at digging deeper into what a political relationship meant to them. Their answers revealed that there were two main definitions of a political network. Some maintained that any member of their political network should agree with their overall political orientation and goals, whereas others only needed to have some working relations with a person, regardless of political compatibility. It was a small sample, but the lines between these two positions were clear: those who prioritised political compatibility were the *reformasi* political elite—people who organised around political ideas and issues. Those to whom political compatibility did not matter were older elites, used to involvement in a type of personalised power machination, divorced from ideology.

The results of the interviews showed that the average percentage of the computational network that elites themselves considered to be part of their political network was 46%, although this figure hides a wide range of different results for each elite, from 20% to 76%. The elites also provided additional information on the context of their relations with people in the computational network, so it is possible to further understand the nature of the reported relations (see Figures 3 and 4).

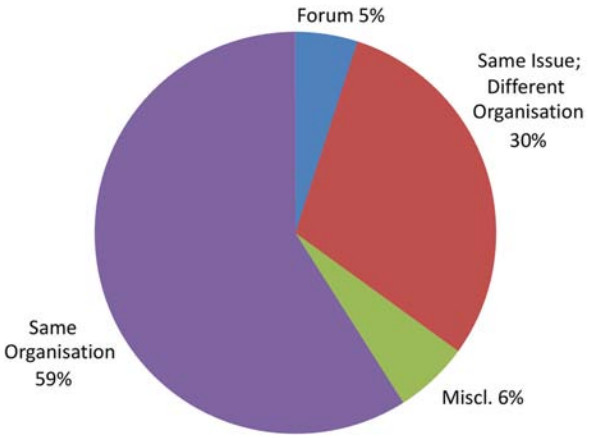

FIGURE 3 *Types of relations in the computational networks confirmed by elites*

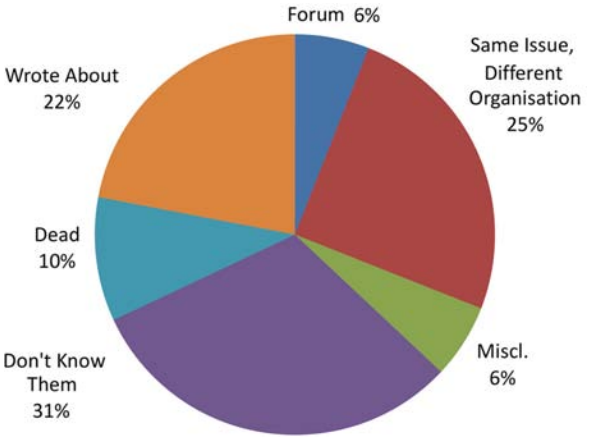

FIGURE 4 *Types of relations in the computational networks NOT confirmed by elites*

Figures 3 and 4 show that being part of the same organisation is a relatively good proxy indicator for being in an elite's political network. It also shows that those who engage on the same issue but are from different organisations are more or less equally likely to be considered by the elites themselves to be part of their network, as is their appearance together on a forum. "False positives"—people who are recorded as being in someone's network using the computational techniques but are not confirmed by the elites themselves—come from categories that are related to the nature of newspaper coverage. A fairly large percentage of false positives (22%) were people who were written about by

the interviewees, so for example, the anti-corruption activist writes about corruption suspects who he did not consider to be part of his political network. Historical relations where elites are mentioned in relation to each other's past affiliations is also a problem in the false positives, shown by the 10% who were no longer living during the time period covered.

Returning to the idea that the elites' own perceptions of their network is in itself problematic, an understanding of their backgrounds could explain some of the results. For example, one interviewee, Sofyan Wanandi, is a top ethnic Chinese businessman who had been involved in advocating business interests to the government for decades. Yet, he maintained that he had no political network and gave me only the names of other business people. This is understandable considering the sensitivity of political-business relations, particularly for the ethnic Chinese. Another respondent, Agus Widjojo, served on the board of the East Timor Truth Commission and saw himself as a mediator between critical civil society activists and the military. His confirmations that he had human rights activists within his political network reflects this. Another, Akbar Tandjung, has been at the centre of Indonesian politics for decades and as an archetypal political operator, was keen to detail his relations with many of the country's top elite.

This is a reflection of what Flyvbjerg describes as "a critical difference between natural and social sciences: The former studies physical objects, while the latter studies self-reflecting humans and must therefore take account of changes in the interpretations of the objects of study. Stated in another way, in social science, the object is a subject" (2001: 32).

Nevertheless, accepting all these difficulties, a valuable insight from the field trip interviews was that the computational network had *some* validity. Considering it was based on co-occurrence only, the figure of 46% was in fact higher than we expected, showing that the basic assumption of some sort of real-life relationship was not an unreasonable one. Figuring out how to automatically discard the "false positives", as well as how to classify the evident variety of confirmed relationships, however, is a much more complex task upon which the remainder of the project will now have to focus.

Conclusion

The ENS project has big ambitions. Working in a challenging multi-disciplinary environment, it aims to develop tools to scale up the conventional techniques of social scientists to make use of some of the digitised texts now available. How successful have these tools so far proven?

The *identification* of elites from the newspaper articles works relatively well, with all of the Indonesia area specialists on the project recognising the groups of people it turns up as political elites. After decades of research on ways to identify populations of political elite, it is clear that there is not a group of objectively verifiable political elites waiting to be uncovered if only the right method could found. Rather, the best we can do is compare alternative methods by exposing their underlying assumptions to better understand what sorts of elites they are each likely to find. It is within this context that the computational method should be considered.

Using co-occurrence to represent elite *relations* is currently more problematic as it identifies a wide range of different relations that cannot yet be automatically distinguished, as well as many "false positives". However, as the interviews with political elites showed, co-occurrence has enough merit to act as a kind of skeleton structure upon which further research can build.

In order to further progress this type of research, social scientists need to close the gap between the nuanced, multifaceted and ever-shifting relations that fill the social science literature and the simpler, more focused and targeted types of relations that are likely to be found using the computational methods. An attempt to do so in this paper has led to an initial conclusion to avoid overburdening the computational techniques with looking for complex relations from the outset and instead focus on capturing one dimension of a relationship at a time.

This approach recognises and works with the strengths of the computational techniques, rather than becoming mired in all of its many limitations. For example, with the elite interviews indicating that co-occurrence is best at uncovering networks of people who work in the same organisation, the best strategy could be to work on obtaining a higher level of accuracy for that type of relation only. It may not be among the most profound insights that a social scientist could hope for on political elites, but it would be no small step to be able to say something about the rate of changeover among the elites of different, high profile organisations over the years.

As noted by some of the other authors in this special edition, working with computational techniques on social science questions is not a simple task. This is contrary to some of the more populist views of "Big Data" as heralding an era where we can just throw all the data we can find into a computer programme and see what results. In practice, it demands a continual effort to check its validity, make connections with other types of information, and filter it in ways that make sense for the types of questions social scientists are interested in answering.

References

Abente, Diego (2009) "Paraguay: The Unravelling of One- Party Rule". *Journal of Democracy* 20(1): 143–156.

Adler, E. Scott and John Wilkerson (2011) "The Congressional Bills Project". Available at: http://www.congressionalbills.org.

Balahur, Alexandra and Ralf Steinberger (2009) "Rethinking Opinion Mining in News: From Theory to Practice and Back". In the Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis, Satellite to CAEPIA 2009.

Burton, Michael and John Higley (1987) "Elite Settlement". *American Sociological Review* 52(2): 295–317.

Dahl, Robert A. (1958) "A Critique of the Ruling Elite Model". *American Political Science Review* 52: 464.

Flyvbjerg, Brent (2001) "Making Social Science Matter: Why Social Inquiry Fails and How It can Succeed Again". Cambridge: Cambridge University Press.

Gooding, Paul (2013) "Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods". *Literary and Linguistic Computing* 28(3): 425–431.

Hicks, Jacqueline, with Ridho Reinanda and Vincent Traag (in press) "Old Questions, New Techniques: A Research Note on Identifying Political Elites Computationally". *Comparative Sociology*.

Higley, John (2009) "Chapter 9. Elite Theory and Elites", in Kevin Leicht and Craig Jenkins (eds.) *Handbook of Politics: State and Society in Global Perspective*. New York: Springer.

Margetts, Helen Z., Peter John, Scott A. Hale and Stephane Reissfelder (2013) "Leadership without Leaders? Starters and Followers in Online Collective Action". *Political Studies* doi: 10.1111/1467–9248.12075.

Martin, Patrick (2012) "Analyzing the First Presidential Debate". Available at: http://dexvis.wordpress.com/2012/10/06/analyzing-the-1st-presidential-debate/.

O'Donnell, Guillermo and Philip Shmitter (1986) *Transitions from Authoritarian Rule*. Baltimore: John Hopkins University Press.

Phithakkitnukoon, Santi, Zbigniew Smoreda and Patrick Olivier (2012) "Socio-geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data". *PloS One* 7(6): e39253.

Procter, Rob, Farida Vis and Alex Voss (2013) "Reading the Riots on Twitter: Methodological Innovation for the Analysis of Big Data". *International Journal of Social Research Methodology* 16(3): 197–214.

Putnam, Robert. D. (1976) *The Comparative Study of Political Elites*. Englewood Cliffs, N.J.: Prentice-Hall.

Reinanda, Ridho (2014) "Towards Extracting a Semantic Graph of Entities from Unstructured Text". Presentation, Benelux Digital Humanities.

Scott, John (2003) "Transformations in the British Economic Elite". *Comparative Sociology* 2(1): 155–173.

Spirling, Arthur (2012) "US Treaty-making with American Indians". *American Journal of Political Science* 56(1): 84–97.

Sudhahar, Saatviga, Thomas Lansdall-Welfare, Ilias Flaounas and Nello Cristianini (2012) "Quantitative Narrative Analysis of US Elections in International News Media". Available at: http://ipp.oii.ox.ac.uk/sites/ipp.oii.ox.ac.uk/files/documents/IPP2012%20Paper-Quantitative%20Narrative%20Analysis%20of%20US%20Elections%20in%20International%20News%20Media.pdf (http://electionwatch.enm.bris.ac.uk/US-Elections-2012/index.html).

Zavala, Dominica (2013) "Disentangling the Fall of a Dominant-Hegemonic Party Rule: The Case of Paraguay and its Transition to a Competitive Electoral Democracy". LSE International Development Working Papers (13)44.