The promise and perils of using big data in the study of corporate networks: problems, diagnostics and fixes

Heemskerk, E.; Young, K.; Takes, F.W.; Cronin, B.; Garcia-Bernardo, J.; Henriksen, L.F.; Kindred Winecoff , W.; Popov, V.; Laurin-Lamothe , A.

DOI
10.1111/glob.12183

Publication date
2018

Document Version
Final published version

Published in
Global Networks

The promise and perils of using big data in the study of corporate networks: problems, diagnostics and fixes

EELKE HEEMSKERK,[*] KEVIN YOUNG,[†] FRANK W. TAKES,[*]
BRUCE CRONIN,[‡] JAVIER GARCIA-BERNARDO,[*]
LASSE F. HENRIKSEN,[§] WILLIAM KINDRED WINECOFF,[‖]
VLADIMIR POPOV[‡] AND AUDREY LAURIN-LAMOTHE[¶]

[*]*University of Amsterdam, CORPNET,*
PO Box 15578, Amsterdam 1001 NB, Netherlands
e.m.heemskerk@uva.nl (corresponding author) / takes@uva.nl / garcia@uva.nl

[†]*University of Massachusetts Amherst, Political Science,*
Amherst, Massachusetts 01003, USA
keviny@umass.edu

[‡]*University of Greenwich, London, UK*
C.B.Cronin@greenwich.ac.uk / V.Popov@greenwich.ac.uk

[§]*Copenhagen Business School, Copenhagen, Denmark*
lfh.dbp@cbs.dk

[‖]*Indiana University Bloomington, Bloomington, Indiana, USA*
wkwineco@indiana.edu

[¶]*Université du Quebec en Outaouais, Gatineau (Québec), Canada*
audreylaurinlamothe@gmail.com

Abstract *Network data on connections between corporate actors and entities – for instance through co-ownership ties or elite social networks – are increasingly available to researchers interested in probing the many important questions related to the study of modern capitalism. Given the analytical challenges associated with the nature of the subject matter, variable data quality and other problems associated with currently available data on this scale, we discuss the promise and perils of using big corporate network data (BCND). We propose a standard procedure for helping researchers deal with BCND problems. While acknowledging that different research questions require different approaches to data quality, we offer a schematic platform that researchers can follow to make informed and intelligent decisions about BCND issues and address these through a specific work-flow procedure. For each step in this procedure, we provide a set of best practices for how to identify, resolve and minimize the BCND problems that arise.*

Keywords BIG CORPORATE NETWORK DATA, BIG DATA, CORPORATE NETWORKS, DIAGNOSTICS, NETWORK DATA QUALITY

Global Networks 18, 1 (2018) 3–32. ISSN 1470–2266.

The age of big corporate network data

Corporations are tightly embedded in networks of power and control and they share board members (which create interlocking directorates), owners and holdings. A sizeable literature has established that these networks facilitate the spread from board to board of corporate governance routines and practices through imitation and learning (Davis 1991; Haunschild 1993; Rao and Sivakumar 1999; Tuschke et al. 2014). As a communication structure, the network promotes the reproduction of existing beliefs and ideas, as well as the dissemination of new ones (Burris 2005; Carroll et al. 2010; Mizruchi 1989). These networks have long formed distinct national business communities and have been part of the organization of national economies. Increasingly, however, they are transcending the national level and forming a new complex global system of corporate ownership and control (Heemskerk and Takes 2016; Heemskerk et al. 2016; Starrs 2013; Vitali et al. 2011).

This fundamental reorganization of contemporary networks of corporate control has coincided with remarkable innovations in research practices. Over the last two decades, the fields of computer science, physics and complexity studies have become increasingly interested in complex network analysis, which has led to numerous breakthroughs in biology, sociology, finance and economics (Barabási and Albert 1999; Battiston et al. 2016; Borgatti et al. 2009; Schweitzer et al. 2009). At the same time, new datasets are now available that allow us to start investigating standardized information on millions of firms and connections between them. Just a few years ago, scholars were manually identifying the 'top 100' or 'top 500' global firms from lists such as the Fortune 500 with a view to evaluating the status of transnational elite ties (Carroll and Sapinsky 2010; Cronin 2012; Davis et al. 2003; Murray 2014). Studies of elite network community structures in certain regions, such as Europe, tended to include a few dozen (van der Pijl et al. 2011), or a few hundred (Carroll et al. 2010; Heemskerk et al. 2013) large firms. Now, scholars have begun to scale up their analyses to global levels by composing lists, for example, of the one million largest firms in the world (Heemskerk and Takes 2016) or the 0.6 million most significant transnational corporations in a structure of global corporate control reduced from 30 million available firms (Vitali et al. 2011). We call this scaling up the emergence of big corporate network data (BCND).

This means that we can now combine advanced analytical and computational tools for analysing big data on the one hand with theories on the architecture of the global economic order on the other. Such studies are likely to proliferate in the years to come, raising new possibilities for research and new questions about the structure of contemporary capitalism (Compston 2013). Complete, or quasi-complete, population studies are particularly promising for network analysts because datasets based on sampling limit the range of techniques and measures that one can soundly apply when conducting network analysis (Marsden 1990; for the debate on sampling issues in interlock studies, see Carroll and Fennema 2004; Kentor and Jang 2004, 2006). More fundamentally, large-scale network data finally hold the promise of overcoming the nagging boundary problem of network analysis. As Allen (1974: 396) stated in his pioneering work, 'the

most satisfactory sampling design for structural analysis is a saturation sample of the entire universe or population; however, this alternative is clearly not feasible for large social structures.' Forty years later, we can confidently say that we have reached the phase where we can use big data to study the entire universe of interest.

Big data, big problems?

While big data bring great promise, they can also bring big problems. Discussions about big data sometimes suggest that their sheer volume should reduce data quality worries (Mayer-Schönberger and Cukier 2013), a train of thought that assumes that it is possible to wash away missing observations and marginal inaccuracies as error. While this is hardly correct for any kind of data – data are rarely missing completely at random – it is a particularly dangerous assumption to make in the context of network-relational data. That is because such missing data can significantly transform network topologies and thus observed network analysis results (Borgatti et al. 2006; Mestres et al. 2008). Some network analytic measures and techniques are robust enough to handle reliably a few missing nodes or edges, but others, and often the more interesting ones, are highly fragile when faced with data incompleteness and sampling bias (Costenbader and Valente 2003).

At the same time, there is a misunderstanding that the central challenge associated with big data, and potentially with big corporate network data, is only that of devising new computing architectures and algorithms (Jagadish 2015). It fuels the widespread perception that dealing with big data simply means scaling up computational capacities and developing new algorithms (Agrawal et al. 2014). We see the challenge of big corporate network data as presenting a set of *analytical* problems, and not simply technical ones. This is not to say that the volume does not change the researcher's relationship with the data. It does, and in significant ways. Utilizing big corporate network data sources from off-the-shelf information providers such as *Orbis*, *Boardex* or *Thompson One* essentially outsources the data collection. Whereas the manual hand-coding of the past was laborious, it provided the researcher with a good grounding knowledge of the data. This intimate understanding of the data is now gone. This leads to a regular confrontation with BCND issues.

Our aim in this article is therefore not to present one specific technical fix, but rather to make a *meta-methodological intervention*. It represents the accretion of efforts from an international consortium of scholars from twelve universities in six different countries. We came together after many bilateral conversations about how to address data quality in the context of the study of corporate elites. When searching for novel practice standards with our colleagues (for example, what to do with missing data in the context of corporate elite connections, or how to report entity resolution issues), we found that we could not find any such standards of best practices. Based on our shared experience of dealing with BCND, we propose a standard process for what we consider to be the most appropriate way researchers should deal with BCND problems, acknowledging that different research questions require different approaches to data quality. There is an urgent need for such standards so that scholars can more effectively measure what they seek to measure, compare alternative data sources and, ultimately, better

accumulate valuable knowledge about what corporate networks look like and how they may be changing. For these reasons, it is imperative to begin a conversation about research process standards *now* if we are to advance the quality of the research community in the future.

In what follows below, we begin by sketching the problems that come with big corporate networks. We put forward a framework whereby we first separate the most fundamental issues with BCND to suggest a structured way to diagnose and fix these issues subsequently, using well-known characterizations. This takes the form of a schematic platform for making informed and intelligent decisions about BCND issues. These occur on multiple levels and involve different iterative steps, and thus we lay out a set of work-flow procedures that researchers can follow to address these issues through a decision tree. Within each level of the decision tree, we provide a set of best practices for how to identify, resolve, or minimize BCND problems that arise. This means that while we suggest methods to reduce uncertainty and noise from the data, our main goal is to assess the extent to which data quality issues exist and what it means for the meaning that we derive from the analysis of concern. We introduce new tools and techniques to diagnose the severity of BCND problems as well as specific techniques and fixes to deal with these problems.

We intend this article not only to help researchers working on existing projects that confront BCND problems but also to encourage future scholars to engage in these data quality issues head on through a systematic process rather than to minimize them. While scholars can and will adapt the specifics of our recommendations to different circumstances in future research, we also hope that reviewers of research use some of the insights we offer here to help improve the peer review process and in the interest of better science.

We do not take a position in the debate on the merits of data driven versus theory driven research, for we believe that the problems we discuss here are relevant for researchers in both domains. We also do not intend our intervention to be specific to the study of corporate interlocks, though we use it as an important running example.

We believe that our suggestions extend wider than this kind of analysis, for they incorporate networks among corporations in general. While many existing studies have examined board interlocks among firms, recent analyses have extended to financial flows across firms (Battiston et al. 2016; Squartini et al. 2013), ties of ownership (Fichtner et al. 2017; Garcia-Bernardo et al. 2017; Haberly and Wojcik 2015; Vitali et al. 2011), and other connections among elite interlocutors of firms that do not constitute board interlocks (Kim et al. 2015). More generally, we acknowledge that the issues we encounter are paramount in other fields of enquiry related to network analysis as well. The suggested diagnostics and fixes may be applicable to these domains.

Big corporate network data: characteristics and issues

The characteristics of big data are traditionally seen through the prism of 'three Vs' – *volume*, *velocity* and *variety* (Laney 2001). More recently, some scholars suggested additional Vs, including *veracity* (Ward 2013), and *variability* (Fan and Bifet 2012).

These Vs provide us with a categorical context we can use to dissect the issues and problems into which we run when working with BCND. In this section, we therefore explore BCND through the lenses of these Vs to determine what issues we need to address.

While *volume* – indicating the sheer amount of data now available to researchers – is the most well-known characteristic of big data, we argue here that the volume *per se* is not problematic in the case of BCND. A typical concern with the *volume* of big data deals with the information processing challenges associated with data analytics (Fisher et. al. 2012). We do not focus on these technical issues because we see it as a misperception that the integration between big data and social science is about technical capacities. Certainly, within the context of BCND the volume is larger than before, but manageable with current tools and techniques. However, the sheer *volume* of the data alters the researcher's relationship to the data, which in turn leads to several (analytical) issues related to the other Vs.

First, BCND feature a *variety* of information. To store information, researchers use distinct types of structured data that generally lack universally employed unique identifiers. While the richness of these data is an asset, different data sources – or even the same data source at different points in time – may not use the same rules for collecting and coding data. One of the key challenges confronting the study of large corporate networks is therefore *entity resolution* – the process of determining whether similarly named firms or similarly named individuals are the same or different actors. In addition, *variety* means that data comparability and completeness may not be consistent across sets of data or different time points. Thus, it is increasingly important to know what mechanisms, in addition to the data-generating process, people use to collect, clean and store the data. Yet, the providers of privileged information are not always keen to share this information. Another key challenge of BCND is therefore to assess the completeness of the data.

Second, *velocity* characterizes BCND. Traditionally, velocity refers to the fact that the flow of data is, apart from being massive, continuous and constantly flowing in from various sources. BCND source databases are updated almost continuously, so the data change quickly with the addition of extra information over time. This leads to new research opportunities, for instance utilizing longitudinal information. It also means that some parts of the database may be updated while others are not. In the case of BCND, we typically see that the more developed and the richer the countries are, the better their corporate registries and hence the higher the *velocity* of the data. This higher *velocity* in some countries compared with others can lead to incorrect comparisons. In other words, the *velocity* of BCND leads to the issue of *accuracy.*

Third, *veracity* refers to the fact that the quality of data is often unclear. For instance, is the information on board composition correct and up-to-date? This relates to the issue of data provenance, which refers to the description of the origin, creation and propagation process of data collecting (Glavic 2014) and the general logic of its extension and priorities. Data are collected through a variety of means and, typically, the precise collection protocols are not transparent and cannot be thoroughly audited. *Veracity* of BCND thus also leads to concerns about *accuracy* and *completeness*.

Finally, *variability* refers to the fact that the way in which the user wants to interpret the data may change over time or according to research question. For example, in inter-firm networks we may sometimes be interested in studying firms with different corporate entities as one entity, whereas if we are primarily interested in the corporate structure we should keep all the firm's legal entities distinct. *Variability* in the use of data requires us to understand how the data are constructed. However, because of the *variability* of BCND, it is crucial that the researcher is clear about its *unit of analysis.* What is it that you want to study? While this is obviously true for all studies, we argue that with big data in general, and BCND in particular, there is an increased risk of errors because data collection is not tailored to the research question. In practice, we often see that researchers devise research questions that try to utilize the full potential of new data sources. This is not in itself problematic, but it means that researchers may be tempted to use units or fields in the data structure as objects of research. This can hold for both the nodes and the edges in the considered corporate network. It is therefore imperative to consider carefully if the BCND that is available does indeed correspond with the proper *unit of analysis*.

Some researchers consider validity yet another V of big data, referring to whether the type of data considered is suitable for measuring the phenomenon in question. For example, in the board interlock network, edges are often assumed to facilitate potential information exchange. Although we may be confident that the board interlock network correctly models the actual board composition, we do not necessarily know about the precise information exchange between the boards on a case-by-case basis. Also, different countries have different governance structures, rules and regulations. A non-executive director in China is not the same as a non-executive in the UK. A big data approach easily allows for the study of, for instance, board interlocks across the globe, but decontextualizing boards and firms may lead to invalid conclusions. One way of seeing this is that validity refers to the veracity, not of the data *per se*, but rather of the researcher's interpretation of the data (such as an edge) as a proxy measure for something else (information exchange).[1] Therefore, it is essential that the researcher has a firm understanding of the theoretically informed unit of analysis. Given that potential problems, diagnostics and fixes for validity are like those of veracity and variability, we do not consider it separately in this article.

Exploring the characteristics of big corporate network data brings us to four basic problems (see Table 1), which are not only relevant for big corporate network data. However, we argue that all studies that use BCND should carefully consider each of these questions. Are you clear about the appropriate unit of analysis? Is there entity ambiguity in your data? How complete are the data? How accurate are the data? These four questions may appear simplistic. However, reviewing the literature we find that typically studies do not report (sufficiently) on these issues. In part, this may be due to the above-mentioned *idée fixe* that when one uses big data we need not worry much about data quality because their sheer volume will counter the effect of missing or incorrect data values. In part, this lack of transparency on these basic questions may relate to the current deficiency of tools and techniques with which to assess the completeness and accuracy of the huge datasets we now use. To remedy this, we propose

several diagnostic routines and techniques for fixing data problems. These fixes fall into two broad categories: *semantic techniques* try to correct the diagnosed problems by using additional attribute information of the data, while *topological techniques* utilize network properties to assess and increase data quality.

Table 1: The Vs of big data mapped to problems in corporate network analysis.

	Unit of analysis	Entity ambiguity	Completeness	Accuracy
Volume	✓	✓	✓	✓
Variety		✓	✓	
Velocity				✓
Veracity			✓	✓
Variability	✓			

Figure 1 is a schematic overview of these four issues in the form of a decision tree. Proceeding through the latter, an honest answer can often be 'not sure'. Therefore, we also suggest some diagnostics to help the research community answer these questions. We hope that the decision tree and the suggested tools and techniques help researchers using corporate network analysis to answer important questions more systematically. Authors can increase transparency by providing an answer to these questions in their methods sections. The next section continues with a step-by-step discussion of each of these questions, diagnostics and fixes, as illustrated by the decision tree. These steps are sequential for a reason. The question about the unit of analysis determines what kind of data to study and select from a source database, and represents an important conceptual step as one related to diagnosis of data quality. One needs to address entity ambiguity before completeness, for incorrect entity resolution may lead to misleading statistics when one assesses completeness. One should address completeness before accuracy because, given that certain segments of the data may be incomplete, we may wish to reduce the sample size to a complete segment or aspect of the data.

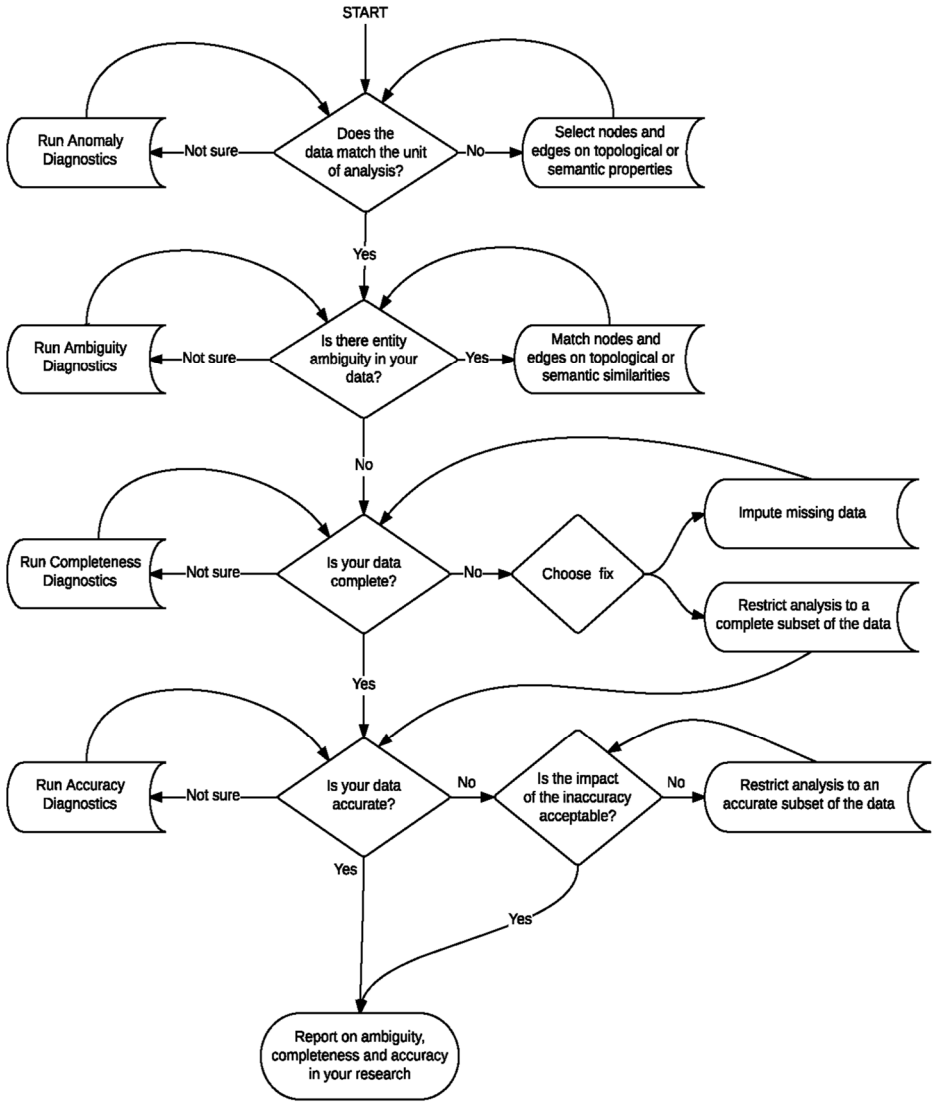
Diagnostics and fixes for big corporate network data

Step 1: Identifying units of analysis

Problems with units of analysis

When we pursue analysis of large-scale networks it can be tempting simply to consider the data within the dataset comprising the network of interest. As students of corporate networks, we must use a meaningful unit of analysis. We must also have a clear definition of what constitutes a firm (node) in a given corporate network of interest and

Figure 1: Decision tree for handling big corporate network data



what constitutes an edge. With BCND this is not always a trivial task since corporations are composed of many interrelated legal entities. As Butts (2009: 416) remarked:

> to represent an empirical phenomenon as a network is a theoretical act. It commits one to assumptions about what is interacting, the nature of that interaction, and the time scale on which that interaction takes place. Such assumptions are not 'free', and indeed they can be wrong. Whether studying protein interactions, sexual networks, or computer systems, the appropriate choice of representation is key to getting the correct result.

When approaching research questions related to corporate network data, one confronts a simple but important ontological question: what is a firm? While this might seem trivial for many kinds of analyses, for the study of corporate networks it is a fundamental question about the definition of nodes and edges. Legal definitions matter because much of the data on corporate networks come from public registers. However, as scholars we may not want to rely on lawyers' definitions of firms. Shell companies, for instance, disturb our common sense about what a firm is. Shell companies are legal entities without any underlying corporate activities, often set up to lower taxes (or, in more malign cases, to avoid corporate responsibility, liability, or to launder money) (Garcia-Bernardo 2017). As such, a board interlock between two shell companies is not theoretically equivalent to an interlock between firms engaged in actual corporate activities (Heemskerk and Takes 2016). Furthermore, shell companies often have boards consisting mainly of lawyers who can have formal board memberships in the hundreds or even thousands. These nodes fundamentally change the network topology in the corporate network of concern and lead to a careful reflection on whether we should consider shell companies as actors in our corporate network. This train of thought essentially feeds back to the initial basic question – what are the nodes and edges in our network and are they commensurable? – and is associated with the boundary specification problem in network research (Carpenter et al. 2012; Laumann et al. 1989).

Whenever we broaden our definition of edges or nodes, our network substantively changes its meaning and function. This is a central issue within network analysis (Butts 2009). Even when the researcher has a clear understanding of what the nodes are, another boundary issue presents itself: what set of nodes and edges are part of the same network? This problem typically emerges when dealing with complete populations of firms in any given geographical context. Here it is advisable to question if one can meaningfully think of any given population as *one network.* If we are interested in studying the Indian or Dutch corporate network, we sometimes want to qualify what comprises this network. Many studies for instance exclude wholly owned subsidiaries of foreign firms; for example, IBM Netherlands is not considered part of the Dutch network (Stokman et al. 1985). With small samples, researchers can hand pick their samples, but this becomes a problem in large-scale databases where we observe huge variations in the kinds of firms.

Unit of analysis diagnostics

There is no single diagnostic for examining if the network data represent the unit of analysis well. We suggest an exploratory approach that takes account of several measures and reflects wisely on the research question of concern. The bottom line is to look for unexpected anomalies in the data. If we are interested in interpersonal networks based on affiliations, producing an appropriate plot of the distribution of affiliations among the population of individuals in the dataset is already likely to reveal anomalies in the data. Distributions of affiliations are highly likely to be long tailed and any obvious spikes at the high end of the distribution could indicate that an identifiable group of outliers is present in the data. Whether we then want to include this group of

individuals or not is an analytical question that should be clarified as we define (or re-define) our unit of analysis. In a similar vein, we can look for deviances from structural characteristics in the data that a certain type of corporate network is known generally to display. If a core–periphery structure is usually found in a particular type of corporate network, but is not so in a set of observed data, this could be caused by a systematic group of outliers that behave strangely (rather than that the actual network of interest does not have a core). If time-stamped data are available, it is possible to look for temporal anomalies. Using the raw data to plot how network-level measures of interest (for example, centralization, cluster coefficients, core-ness) vary over time can be useful here. If measures are volatile in ways that one cannot explain temporally (for example, seasonality), we may want to check if alien groups enter the network of interest and disturb otherwise stable structural features.

Fixes for unit of analysis problems

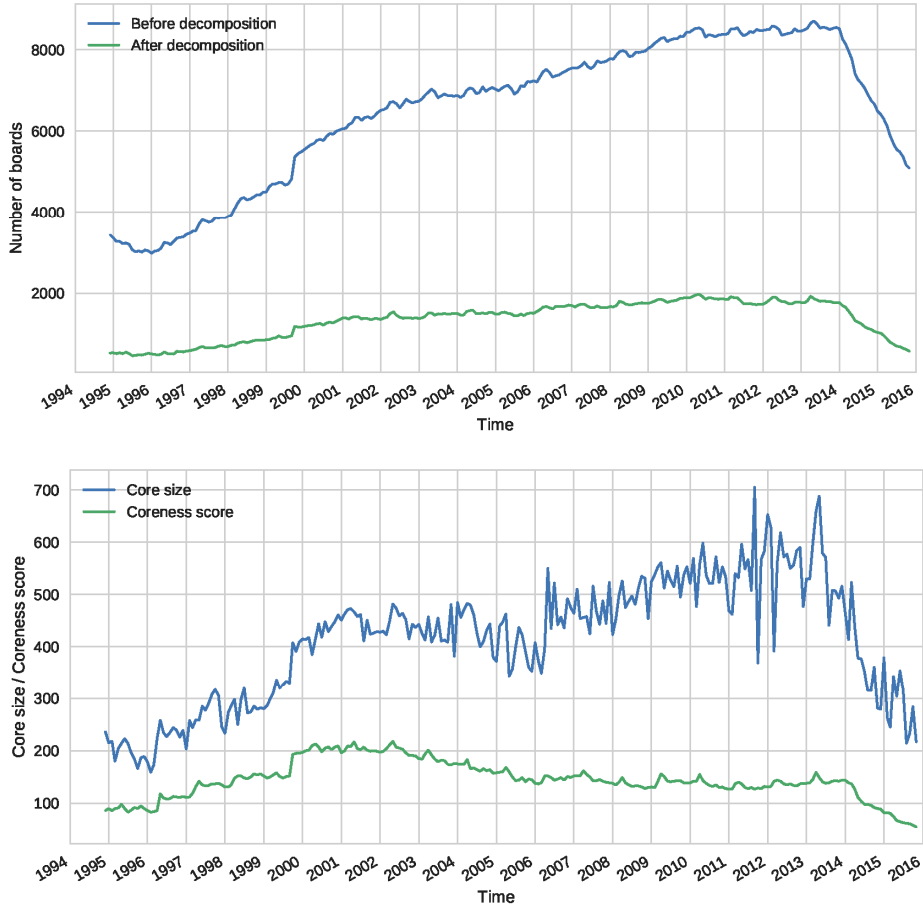
Two main ways of fixing the data problems raised above can be identified. A semantic approach is possible if we can locate a certain type of actor or edge as the root cause of our data anomaly – either from empirical knowledge about the network or from analysis of variance in node or edge attribute data. In that case we can make sense of our problem and make informed decisions about whether to exclude the source of the problem through targeted sampling. This approach is closer to what we might term ordinary data cleaning, regardless of whether search and matching algorithms aid this work or one does it manually. A topological approach by contrast excludes certain nodes from the network of interest based on certain structural characteristics that such nodes display (such as degree), and thus moves towards the more analytical–methodological end of the 'data cleaning and quality assurance' spectrum. We illustrate how a combination of the two approaches can be useful in identifying, and dealing with, data anomalies.

Henriksen et al. (2016) study corporate networks of board members for the complete population of Danish firms 1990–2015. Their data set comprises 422,020 individuals, 208,417 boards and 1,677,688 board memberships with start and end dates of these memberships recorded. Building on Useem's (1984) work on corporate elites, the authors set out to apply dynamic K-core decomposition to understand the temporal evolution of the corporate 'inner circle' in Denmark. K-core decomposition works by recursively pruning nodes with lower degrees and thus successively identifying subgraphs of increasing degree centrality (Batagelj and Zaveršnik 2003). As the threshold for entering the successive subgraphs increases, the subgraph identified becomes ever more cohesive. Based on their detailed spell data they could create monthly time slices of the entire network and apply the same K-core decomposition procedure to all those time slices, in turn figuring out if the size and composition of the core was stable over time. Using this well-established method, it turned out that the composition of the core was highly unstable and its size varied tremendously.

What caused this instability was not, however, a fracturing of an 'inner circle' as found elsewhere (Chu and Davis 2016), but data anomalies such as those described above, where extreme degree values appear because of the presence of shell corporations. The method breaks down because shell companies form their own internal

communities, which are only loosely connected to the true global centre of the network. The degree of nodes within these communities is based on highly redundant ties within heavily overlapping boards. K-core decomposition is ill suited to deal with such situations.

Figure 2: The effects of K-core decomposition on the number of boards (top) and core size/coreness score (bottom), visualized over time



One can correct this situation by introducing path-based centrality measures into the decomposition method. Introducing an additional threshold based on betweenness scores into the pruning process allows one to ignore such locally central K-cores. Insofar as the interesting unit of analysis is a global core in a network, this method deals well with data quality issues such as the presence of shell corporations. Before introducing the betweenness decomposition method, it was not possible to identify any stably convergent core because the highly central board members of shell corporations overly affected the coreness thresholds. After the introduction of betweenness into the pruning process, a stable core emerges as seen in Figure 2.

Identifying the problem, discovering why it was a problem and knowing how to fix it required an exploratory use of both the semantic and topological approaches, where defining the unit of analysis and the population of a network is part of the process of analytical discovery, relying in part on familiarity with network analytic tools to understand topological characteristics and in part on more simple methods of finding data anomalies such as sampling and checking semantics.

Step 2: Examining entity ambiguity

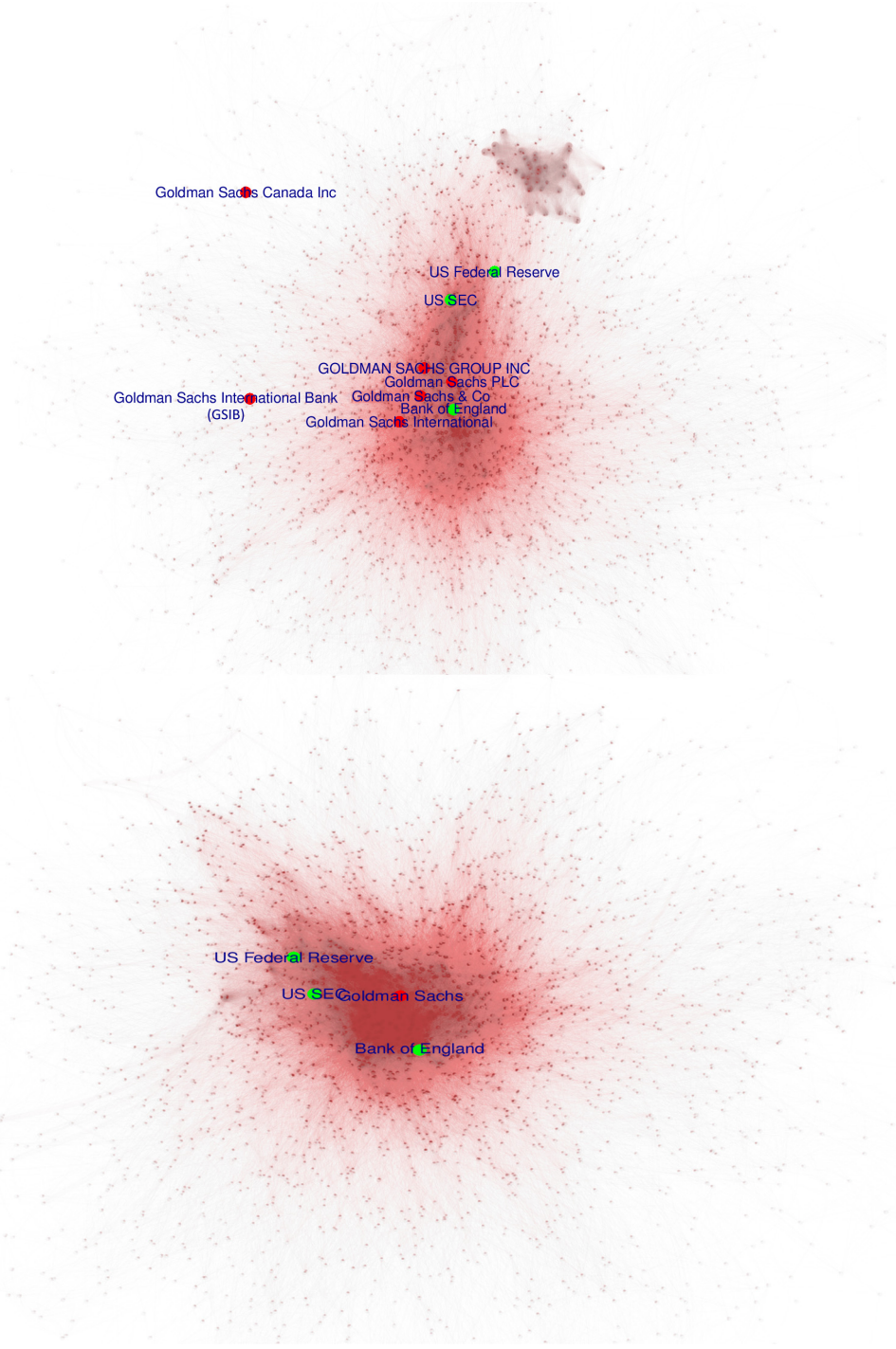
What is the problem of entity ambiguity?

A low quality and integrity of corporate network data poses a fundamental threat to the validity of inferences drawn with big data. A simple example, which utilizes data from *Boardex*, illustrates this point. To investigate connections between public authorities and large global firms, researchers took the first-and-second degree board interlock connections from just three significant financial regulatory authorities in the North Atlantic – the Bank of England, the US Federal Reserve Board, and the US Securities and Exchange Commission. We highlighted these public entities within the network in green in top part of Figure 3. Also highlighted, however, is 'Goldman Sachs'. Yet, as one can see, there is not one Goldman Sachs but five. The centrality of Goldman Sachs in this network is unknown; if one wanted to know the connections between Goldman Sachs and these selected public entities, not only would there be clear biases in the data but there would be five different measures generated for each. This kind of problem with entity resolution will bias measures of network structure, and the problem will only grow more severe with the expansion in size of the network. In the context of traditional datasets of a few dozen or hundred firms in a network this may be an insignificant problem, for manual checking or sorting can resolve duplicate entities efficiently and comprehensively. In a big data context, it is unfeasible to do this comprehensively. The lower part of Figure 3 shows an example of a 'resolved' network (see Marple et al. 2017), in which not only Goldman Sachs but many other entities in the network have been resolved, generating significant changes in network structure and revealing the more genuine location of Goldman Sachs within the network.

One keystone form of entity resolution in a big data context is to use string matching algorithms. String matching algorithms are essentially a procedure in which one compares all entries in the data with all other entries in the data, and this computes a similarity score for each pair of data entries. Then, one can deem the most similar pairs (namely, with the highest score) identical and subject to replacement. In the above example, one can reduce the two Goldman Sachs's to one, and rewire the subsequent network to ensure greater accuracy. Scholars have a variety of string matching algorithms at their disposal (discussed below), which often entail measures of similarity across entity names. Yet, even string matching only works with a degree of confidence; given the absence of manual checking the confidence intervals being relied on need to be transparent (see Garcia-Bernardo and Takes 2016).

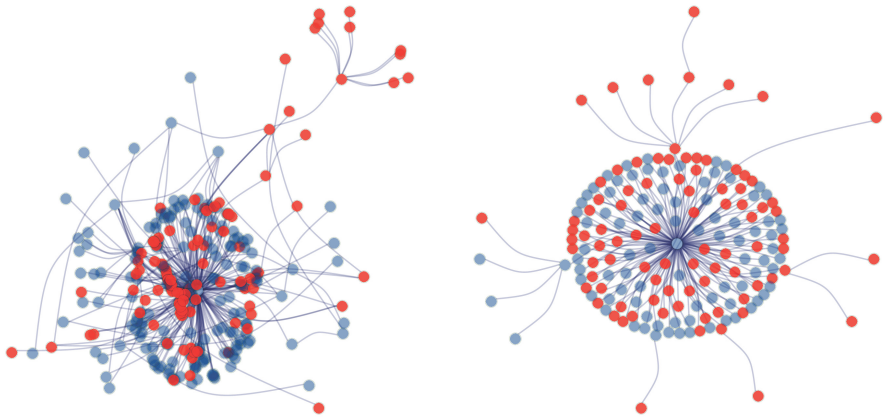
Yet, not all entity resolution issues are this simple. Many firms are part of highly complex corporate ownership structures that compound entity resolution challenges.

–Figure 3: Goldman Sachs within ego networks of the Bank of England, US Securities and Exchange Commission (SEC) and US Federal Reserve Board, original data (top) and resolved (bottom)



According to the *LexisNexis Corporate Affiliations* database, for example, 229 different legal entities, including subsidiaries and shell companies, make up the Bank of America. The nature of such corporate hierarchies is likely to generate biases in network structure if left uncorrected. To illustrate the complexity of corporate forms, Figure 4 provides a network representation of two large global corporations – Citigroup and Exxon Mobil. With the global parent in the centre, from subsidiaries, separate holding companies and shell companies that run through the corporate hierarchy of the parent, one can represent each successive level of the firm. The blue dots in the network represent legal entities that have the name stem of the global parent in their name (for example, 'Citi' or 'Exxon') and that one could potentially resolve through string matching. The red dots however represent legal entities that do not have the name stems of the parent in their name. Exxon Mobil, for example, contains entities in its corporate hierarchy such as 'Houghton Realty Trust', which branch out from 'XTO Energy', which is a subsidiary of Exxon Mobil.

Figure 4: Two corporate hierarchies: Exxon Mobil (left) and Citigroup (right)

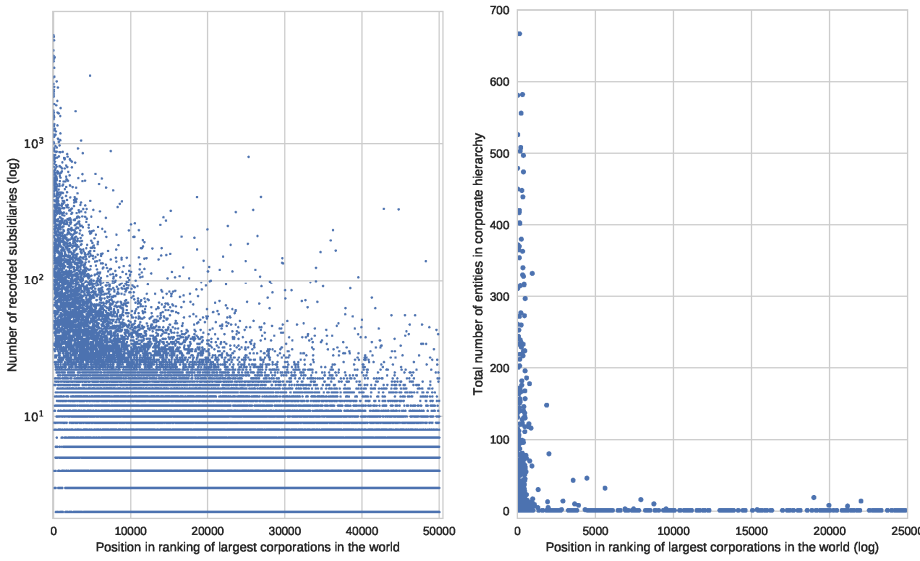


Diagnosing entity ambiguity

We recommend running simple diagnostics if a researcher is unsure of the need to resolve entities. These can come in the form of simple string matching algorithms that, for example, search for all variations of the name 'Met Life' and replace accordingly. Note that this will only identify certain kinds of entity ambiguity issues. As such, it may make sense to gather auxiliary information on the structure of a given corporate hierarchy (such as ownership data) for a firm that is prominent in the data, and then run automated string search algorithms for all firms in a complex corporate hierarchy structure.

Corporate hierarchies not only include subsidiaries of a holding or parent firm, but also specialized holding companies and non-operating entities (otherwise known as shell companies). However, it is possible to diagnose the best way to approach resolving the data given the constraints a researcher faces and given the research questions they are pursuing. Figure 5 (left side) shows a scatterplot of the (ln) number of recorded

Figure 5: Number of subsidiaries related to global parent (left) and total number of entities in corporate hierarchy (right) for the largest corporations in the world

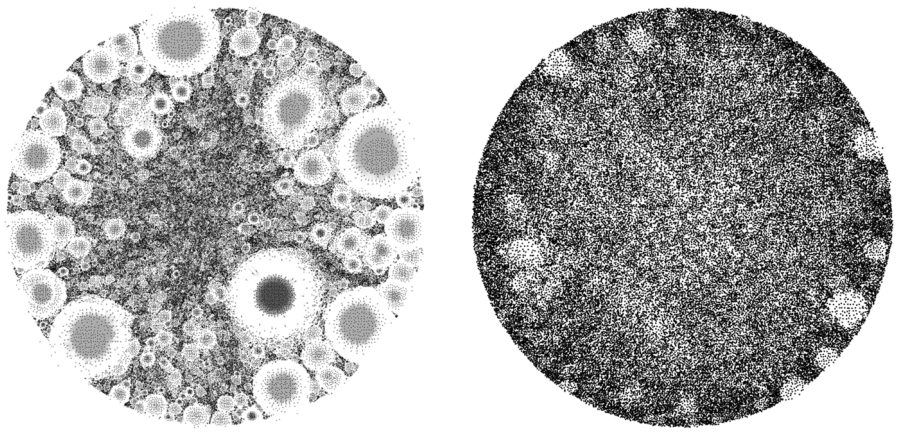


subsidiaries of the largest half million firms in the world, on a global ultimate owner (or GUO) basis, using total assets as the indicator of firm size and using data from *Orbis*. The slope of this relationship reveals that the larger the firm, the more subsidiaries it is likely to have. The right side of Figure 5 shows a more select example with less off-the-shelf data. Because subsidiaries are only one form of entity, we counted the length of corporate hierarchies across the universe of corporations by randomly sampling 100 entities from four different strata of firm size (from ranks 1 to 500, 500–1000, 1000–10,000 and 10,000–25,000) within the largest 25,000 firms across 59 countries from *Orbis*. For each of the 100 firms sampled, we looked up the detailed corporate hierarchy information, which included subsidiaries, branches, units, holding companies, and non-operating entities related to the global parent. For each distinct entity, we counted an additional unit of length in the global parent's corporate hierarchy (Marple et al. 2017). The regularity found within the data, illustrated in Figure 5, is largely the same as for the subsidiary data described above. The larger the firm (ranked in terms of assets), the more entities there were within its corporate hierarchy. In the global distribution of firms, the 'top end' of firms has the longest hierarchies. Such empirical regularities tell us which kinds of firms are likely to have 'longer' corporate hierarchies than others. Thus, when working with BCND research questions, it helps to narrow the range of firms on which to focus in terms of entity ambiguity fixes.

One can also diagnose entity ambiguity through measures of network structure, namely, the topological approach. For example, in an unresolved network of director interlocks among financial firms in the USA, a researcher will quickly find numerous director ties between for instance 'Bank of America NA' and 'Bank of America Securities'. Thus, in some instances entity ambiguity may be identified through

abnormalities in tie structure. One way to identify such abnormalities may be through plotting a diagnostic of edge width across a modelled network. For example, plotting the network using strong attraction between connected nodes and strong repulsion among unconnected nodes can show clusters of highly connected nodes (See Figure 6). These nodes are sometimes local branches of small companies, which one can join together or exclude from the sample. Garcia-Bernardo and Takes (2016) explain this method in detail and, apart from visual inspection, also proposes a number of topological network metrics that can be used to characterize such dense clusters.

Figure 6: Network visualization of the Swedish board interlock network using the ForceAtlas2 algorithm (made in Gephi, http://gephi.org) for the raw network (left) and the same network corrected for entity ambiguity artefacts using the method described in Garcia-Bernardo and Takes (2016) (right)



How to deal with entity ambiguity

To address entity ambiguity problems, the unit of analysis may require merging together related nodes. For example, we may be interested in the relationships between corporations, and thus would like to merge all firms involved in the corporate structure. The best approach here is a topological one in which we use information about ownership to merge related companies. However, we do not always have ownership information. In those cases, we can merge companies that cluster tightly together (topological approach, see Figure 6 above), or use firm names or other firm attributes (semantic approach). The entity ambiguity problem persists when similar company names correspond to different companies (for example, 'ASN' and 'ABN'), and when different names are part of the same corporate structure (for example, 'Zao Master D' and 'Beta Properties INC' are part of 'METLIFE INC'). It is of course possible to utilize this kind of information when it is available. Marple et al. (2017) and Young et al. (2017) utilize large lists of corporate family structures, including branches, subsidiaries and shell companies, among the largest 500 corporations in the world as the basis to batch-replace existing names in the network.

Utilizing string matching, namely merging two nodes if their name is similar or rewiring the network, is quite crucial in all these processes. One can measure similarity by the effort it takes to convert one string into the other by modifying individual characters (edit-based measures), by the number of words or n-grams that are shared between the strings (token-based measures), or by a combination of both (hybrid measures) (Bilenko et al. 2003; Cohen et al. 2003). Since different variants of a company name usually differ in the ending (for example, 'Bank of China limited' versus 'Bank of China LTD'), one prefers algorithms that give lower weights to the end of strings. This is the case in the Jaro-Winkler (edit distance) algorithm (Winkler 1990), and the term frequency–inverse document frequency (TF–IDF) using n-grams (token distance) algorithm (Salton et al. 1975). Metadata information can provide useful supplements for entity matching. If two companies share an address and have similar names, it is likely that they are the same company. Moreover, one can combine several similarity values using machine learning algorithms such as neural networks (Bilenko et al. 2003; Cohen et al. 2003).

Given size and computational capacities for string matching in extremely large adjacency matrices, one can perform string matching on weighty edges above a given reported (high-in-distribution) threshold, where false ties are likely to exist (for example, MetLife Inc. has 200 connections to MetLife) because of the nature of large corporate groups. Another way to exploit network structure as part of an entity ambiguity fix is to utilize community clustering. Marple et al. (2017) use community detection algorithms to separate node names into clusters, which are then sub-processed using string matching methods within each cluster to ensure greater accuracy in name replacement.

Because entity ambiguity problems can be highly complex, researchers might use multiple methods, but in each case, should report entity ambiguity statistics – such as the frequency of name replacements and, if possible, the precision-recall estimates associated with some form of 'ground-truthed' subset of the data. The data that is used for ground-truth will inevitably have to be a sample. If the sample is large enough and the entity ambiguity problem significant enough (or *unknown* enough), one can report the performance of each entity ambiguity fix through forms of measurement developed within the computer science community that measure precision and recall performance of a given method. Precision, or positive predictive value, is the faction of retrieved instances that are relevant, while recall, otherwise known as 'sensitivity', is the fraction of relevant instances retrieved. Synthetic scores exist of precision and recall performance that researchers using BCND data can utilize, such as F1 scores and measuring the area under the ROC (receiving operator characteristic) curve, which can help a researcher decide which entity ambiguity fixes are generating the best performance. An ROC curve is a statistical metric used to visualize the performance of any classifier with two possible outcomes. It represents a plotting of the true positive rate against the false positive rate to understand the tradeoffs between sensitivity and specificity of a given classifier. An F1 score is another way to assess the performance of a classifier, as it measures the combined performance of precision (the fraction of retrieved instances that are relevant) and recall (the fraction of relevant instances that are

retrieved). For an F1 score to be high, both precision and recall should be high. These should be used in the light of what earlier diagnostics suggested, and the potential sensitivity of the network-relational measures that are ultimately being pursued. Treating entity ambiguity seriously and systematically not only facilitates the goal of analytic transparency and hence reproducibility but also will provide other researchers with information about error rates and the severity of entity ambiguity issues for specific research problems with large off-the-shelf big data datasets.

Step 3: Completeness

Problems of data completeness

Data incompleteness can affect the trustworthiness of the inferences made from statistical models of relationships in the data (Rubin 1976). Missing data in a network context is generally more problematic than it is in non-network contexts because of interdependencies in network data (Borgatti et al. 2006; Kossinets 2006). Missing information in networks can have a multiplicative effect: each missing link directly affects two nodes and indirectly potentially affects many others. Completeness affects centrality scores, community detection algorithms and comparisons between networks (Žnidaršič et al. 2012). All these metrics are critical for descriptive or inferential analyses of networks.

There are three main types of data missingness in networks – omission of actors and/or affiliations that exist in the network due to boundary specification; non-responsiveness to surveys used to construct the network data or an inability to construct a full network from observational data; and censoring according to a node-level characteristic such as size or prominence (Kossinets 2006). There are several reasons why some data may be missing, some of which are most relevant for our purposes. First, information providers usually collect complete data for large companies and at least partial data for medium-sized companies, but may not report even the existence of many smaller-sized companies. For example, the database *Orbis* has complete information about Greek companies with more than 250 employees but only contains around 2 per cent of the companies with fewer than ten employees (see Table 2). While it is true that small companies are often less significant in corporate networks than large ones, ignoring the missing data could still significantly alter our results. In addition to differences in filling requirements, data completeness is higher in developed economies than poor economies and tax havens. Second, some data, particularly those describing interdependencies such as ownership relationships or corporate interlocks, may be collected egocentrically, and parts of the network will appear less connected than others because of the sampling procedure.

Completeness diagnostics

Missing information can refer to the nodes themselves (for example, missing companies or people), to the edges (for example, missing director positions or ownership relations) or to metadata (for example, financial information). Moreover, the data can

Table 2: Percentage of companies present in data vs employee range for the corporate data of Greece (comparing Orbis corporate data with OECD statistics)

Number of employees	0–9	10–19	20–49	50–249	>250
Percentage present	1.3	32.4	58.5	79.0	117.5

be missing completely at random (MCAR), or missing with a probability that depends on an observed variable (missing at random, or MAR) or an unobserved variable (missing not at random, or MNAR). Missing metadata usually correlate with another observed variable – for instance it is more likely that we lack data on firm assets if we are also missing data about the revenue of that firm. Because the variables are related, missing metadata can be imputed if some of the metadata are observed. We do not recommend deleting cases under MAR since it produces large biases (Rubin 1987). Missing nodes or edges are commonly correlated with an unobserved variable; this makes them harder to study because imputation is unlikely to reduce bias or inefficiency.

Missing nodes can alter the results significantly. For instance, if we wanted to analyse the characteristics of the agriculture sector using data from Orbis, we would find out that the average Mexican company is larger than the average US one. But this is due only to the better recording of small companies in the USA. Because our results can be erroneous if there are missing nodes, it is paramount to assess the completeness of the data. Completeness diagnosis consists of the comparison of the data (or a subset of the data) to an external database that is known to be complete. Importantly, this step will often require aggregating the data by sector, country, or type of company. Table 2 shows the completeness of the Greek data from the Orbis database by comparing it with Eurostat data. This step provides a first assessment of the type of data that are missing – small companies. If the first step reveals that we have missing data, we need a finer characterization of the missing data. In this second step, we look for the pattern of the missing data. For example, the distribution of most firm economic indicators, such as operating revenue, assets, or number of employees follow lognormal distributions. By comparing our database with external databases, we can characterize the distribution of the missing data.

Although missing edges do not affect network measures as strongly as missing nodes, they can still affect the analysis (Kossinets 2006). Diagnostics also require comparing our data with a complete database. Because complete databases of edges are not readily available, we usually rely on manual checks of a sample of the data. These manual checks in our experience usually show that while big companies have complete information about directors and ownership relationships, small companies lack such information, and thus have more missing edges.

Addressing data completeness

Once we have a clear understanding of the type of missingness in the corporate data we can deal with the problem in two ways – either by restricting our analysis to a part of the network with good-quality data, hoping that the missing part does not bias

inferences taken from the part we observe; or we can seek to improve the quality of the data to mitigate the effects of missingness. If we choose the latter, there are two basic ways to operate under conditions of incomplete data – an approach based on leveraging the assumptions regarding sampling procedures, and an approach based on imputation of estimated data in place of the missing data. If we choose to impute data, we should undergo the process transparently and, if possible, repeatedly.

The simplest missing data to correct are metadata (for example, the attributes of a firm or individual). Unless the amount of missingness is large, one can impute these with normal statistical procedures in the tradition of Little and Rubin (1987), including modern implementations such as multiple imputation and hot-decking (Blackwell et al. 2015a, 2015b; Cranmer and Gill 2013). These provably reduce biases and improve efficiency when data are missing at random.

Non-metadata missingness can exist at the node-level or tie-level. These are more difficult to impute than metadata, but some reasonable strategies have been developed. Information missing randomly at the tie-level (for example, through representative sampling) is the more straightforward of the two, and can be imputed by inference using the latent space positions of nodes to replicate missing edges (Ward et al. 2003). This approach might be useful under conditions of representative sampling or egocentric analysis. Huisman (2009) notes, however, that 'simple' single-imputation methods, namely ad hoc methods that do not involve multiple imputation, are frequently biased. Until very recently this left few options for scholars working with missing network data other than deletion.

The most flexible type of imputation strategy for missingness, either at the node or edge level, involves estimation of missing values using the likelihood-based exponential random graph model (ERGM) family (Cranmer and Desmarais 2011; Handcock and Gile 2010), with extensions for Bayesian 'data augmentation' (Koskinen et al. 2010, 2013). In these models, expected values are imputed for missing data during the estimation of the ERGM parameters; because ERGMs are estimated via simulation (usually employing standard Markov chain Monte Carlo methods), the missing data are imputed many times. While still quite new and rare, these ERGM-based procedures have been successfully implemented on real-world network data suffering from quite complicated patterns of missingness (Desmarais and Cranmer 2012; ,Wang et al. 2016), and have been mathematically extended to temporal ERGMs (Leifeld et al. 2017; Ouzienko and Obradovic 2013). We need more validation to understand fully the properties and reliability of these model-based methods, but they possess considerable promise for scholars analysing network data containing missing information.

We know certain that data missingness presents serious problems in a network context even if the data are missing at random. Scholars should correct bias that could emerge from missingness or, at least, understand its likely effects.

Step 4: Accuracy

Problems of data accuracy

In general, accuracy refers to whether a measurement of data conforms to the real world. Specifically, we are concerned with whether the data consistently and correctly

match our conceptual understanding of corporate networks. Because corporate data are typically gathered in a rather indirect route via annual reports, business organizations and (intermediary) information providers, and furthermore changes over time (respectively the veracity and velocity aspects of big data), accuracy can be a significant issue. It affects whether the existence of a node or a link is correct and current. Furthermore, accuracy is related to the question of whether the observed data accurately represent the type of network structure we are interested in studying.

An example illustrates how inaccurate corporate data can lead to an incorrect corporate network. We create a board interlock network from data on firms and directors in Panama, modelling social ties between firms that share board members. The data provider (Bureau van Dijk's *Orbis* database) aggregates data from different information providers in different countries, and is thus dependent on these data providers for the quality of its delivered content. In *Orbis* there are 841,487 active firms for Panama. This substantial number suggests that a sizeable proportion of the firms in Panama have been collected. To study the Panama board interlock network, for each of the 628,289 firms with actual information on board composition, we selected the senior directors of people listed as currently holding a position at a particular firm. This yielded 3,172,041 unique director positions. In total, there were 1,207,541 unique directors. Given the considerable number of board interlocks and directors, we may feel that we are on the right track to extract a sensible network based on interlocking directorates. When we further inspect this data, for example by looking at the average board size – which is 3,172,041 / 628,289 = 5.05 – the data still seem accurate. Even the distribution of the number of directors per board seems sound, as shown in Figure 8, with an average of around five and a few far less frequent larger boards (note the logarithmic vertical axis). When we examine the average number of positions held by a director – which is 3,172,041 / 1,207,541 = 2.62 positions – there is still no reason for alarm. In fact, it suggests an exciting number of interlocks.

However, when we look at the distribution of the number of positions held by a director in Figure 7, we see something alarming: in Panama, there are directors with extremely large numbers of positions, led by one director in our data holding 16,744 positions at different firms. The names of these directors in the tail of the distribution are not common names that name matching or entity resolution software have wrongly matched by, as the majority appear to be actual unique names of directors with an enormous number of positions.

Including these directors is clearly not beneficial to studying the board interlock network for at least two reasons. First, from a theoretical (semantic) point of view it is unlikely that this person facilitates many interlocks with clear causes and consequences for Panama's corporate structure. From a network-topological point of view, if we study the firm-by-firm network of Panama, then this director alone would create (16,744 * 16,743) / 2 = 140,172,396 interlocks as part of a fully connected clique of firms. Clearly, this is beyond the scale of any meaningful corporate network, especially given that there are almost 100 directors with that many positions.

Second, we should note here that the projection from the two-mode board-director network to the one-mode board interlock network is responsible for the quadratic increase

Figure 7: Distribution of the board size in Panama

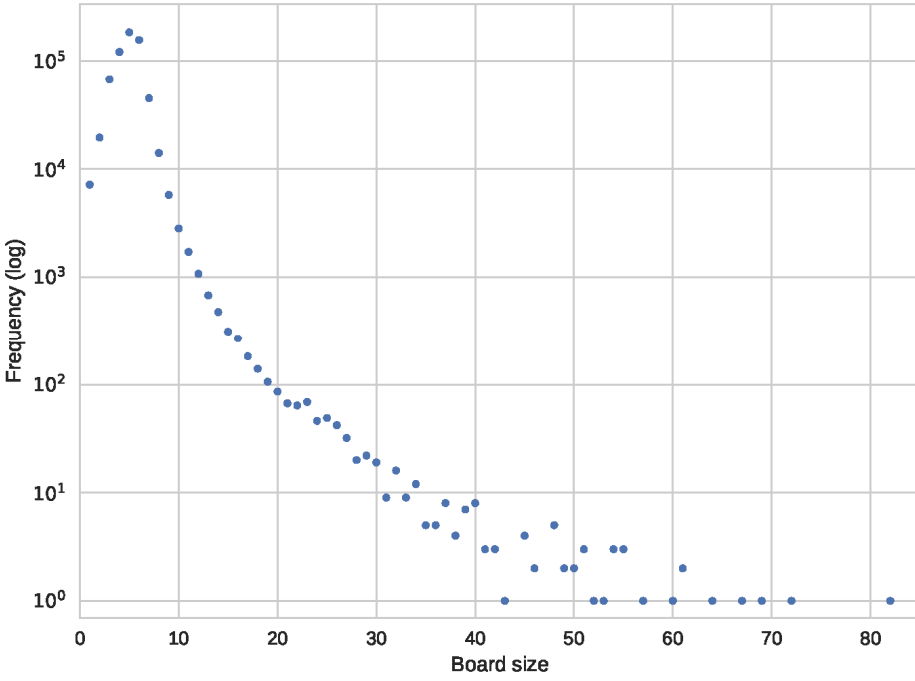
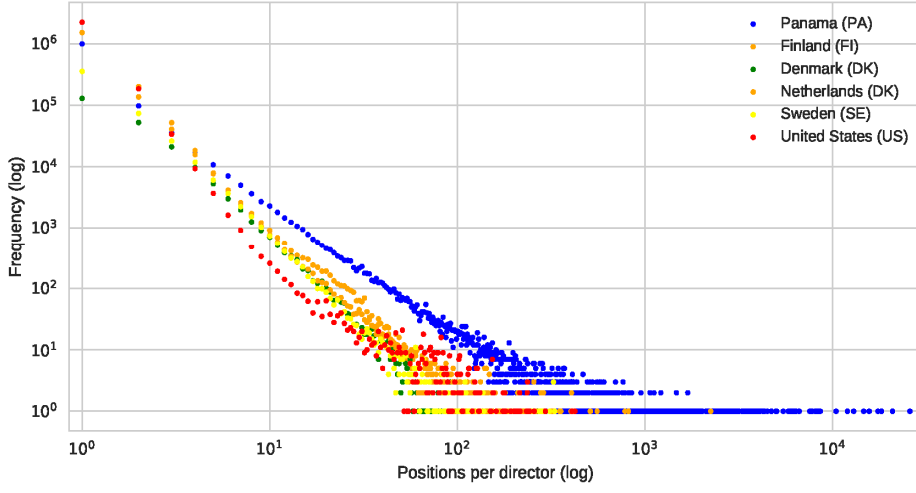


Figure 8: Distribution of the number of director positions in five different countries



in the number of interlocks, a general problem that arises when projecting two-mode networks with skewed degree distributions (Neal 2014). Yet, most likely, the example above is a by-product of Panama's now well-publicized status as a tax haven and host to large numbers of shell companies established for evasion purposes. Indeed, when we

compare the distribution of the number of positions per director with other countries, such as Finland, Denmark, the United States, Netherlands and Sweden, as shown in Figure 8, the long tail in the Panama distribution contrasts markedly with other states.

Accuracy diagnostics

One obvious way of detecting the issues described above is by doing an analysis of the distribution of the data points. Looking beyond mere totals and averages, searching for outliers in terms of frequency and value (thus, looking at the vertical and horizontal extremes and the outliers of this distribution) may allow one to catch errors or problematic idiosyncrasies in the data rather easily. More specifically, if we were generating the global board interlock network, we would generate these distributions for each of the natural groupings of firms into countries, and see if the distribution for each country makes sense. Ultimately, without manually inspecting the millions of director positions, we can automatically obtain insights into outliers. Of course, one can manually look up these outliers in the data to determine the actual reason.

Addressing data accuracy

One way of solving the issues pointed out in the example above is by setting sensible filtering thresholds based on what is known about the data. For example, suppose a researcher finds a given corporate board member with many more ties to other firms than average. In such instances, the researcher should transparently report the decisions taken to exclude or filter data. Investigations found that these exceptional linkers/firms can be particularly elaborate shell company structures (Henriksen et al. 2016). Indeed, one often encounters the presence of such administrative ties (rather than social ties) between firms (Heemskerk and Takes 2016; Takes and Heemskerk 2016). One solution is to take the presence of such administrative ties into account when interpreting the results. A second is to filter data to exclude such formations. Regardless, one should report the precise extent and steps in every case. This is especially important given that normal possibilities of replication for BCND data are often unavailable given the proprietary nature of the data *and* because BCND data purveyors often update their bespoke datasets in real time.

More generally, when data are so large that one cannot assess their quality by hand, we need to place more reliance on quantitative inspection – examination of totals, averages and distributions – rather than qualitative inspection. One can compare these with extant studies of a similar regulatory context and, over time, check for the presence of major discrepancies in data accuracy arising from data volatility. Quantitative inspection also helps to segregate the data and compare the different quantities across different segregations of the data to detect outliers. For example, by segregating the data illustrated in Figure 8 above by country, we find significant outliers in Panama, which we know from the so-called 'Panama Papers' and other sources to be home to a vast number of shell companies and tax-evasion entities. It is worth asking whether the prevalence of shell companies is accurately capturing the concept of interest, which may be corporations that are active in some type of production. If so, then scholars may need to prune the data according to some characteristic – an above-zero number of employees, for instance, or some output-based characteristic – so that the

prevalence of shell companies does not distort the inferences one can make from these data.

There is no panacea to these issues, no perfect statistical 'check' for data accuracy. Correcting data that inaccurately capture the corporate structures of interest requires wisdom and patience, and should therefore be done transparently. Data inaccuracy may also relate to some other issue with BCND – particularly data completeness and entity resolution – so a holistic strategy to ensure that the data are in proper condition to perform the analysis is desirable whenever feasible.

Discussion: toward common goods

In this article, we have sought to add productive fuel to the conversation over data quality when utilizing what we have called big corporate network data (BCND) problems. Even prior to the advent of data on this scale there was a rich discussion regarding how best to study elite networks (Carroll and Fennema 2004; Kentor and Jang 2004, 2006). These and other longstanding issues within this specialized literature – among them what a given edge-relationship actually 'does' (Cronin 2011; Mizruchi 1996) – do not go away. They simply get compounded and added to a litany of other research challenges.

We have advanced a framework to help guide not only individual researchers but also future discussions among the research community when it comes to data quality and the means of addressing it. Researchers should identify whether the data match the unit of analysis; address entity ambiguity, data completeness and data accuracy; and report on these steps and on their preferred diagnostics and fixes. We introduced new tools and techniques to diagnose the severity of BCND problems, as well as specific techniques and 'fixes' to deal with these problems. Specifically, within each level of Figure 1, we provided a set of best practices for identifying, resolving, or minimizing the BCND problems that are known to arise.

Our contention is that the research community would benefit greatly from walking through the flowchart proposed in Figure 1, or something close to it; then, transparently reporting on each step is a good recommendation for forthcoming research about corporate networks. A variety of diagnostic tools for the unit of analysis, entity ambiguity, data completeness and data accuracy, as summarized in Table 3, support each step of the decision tree. As discussed earlier, one can deploy many existing diagnostic tools to improve the rigor of BCND; we recommend extensive and transparent use of these. Yet, while these provide a standard set of metrics that allow interpretation of the validity of the analysis, the context of the metrics remains ambiguous as little is yet known about the typical distributions of these metrics in standard corporate settings. This opens a research agenda for the further development of these diagnostic tools.

This list of suggested diagnostics and fixes will no doubt change and improve over time. While we gear our intervention towards the community of scholars working on corporate networks, we acknowledge that the issues we encounter are paramount in other fields of enquiry related to network analysis as well. The suggested diagnostics and fixes may be applicable to these domains.

Table 3: BCND diagnostic toolkit

Decision Step	Current Diagnostic Tools	Future Development
Unit of Analysis	Degree and edge distributions. Core–periphery analysis. Cohesiveness analysis (for example, centralization, cluster coefficients, coreness). Dynamic k-core decomposition.	Sector, country and company type norms for degree and edge distributions, core–periphery structures and cohesiveness.
Entity Ambiguity	String matching algorithms. Corporate hierarchy length. Utilizing network structure such as edge multiplexity distribution, community clustering.	Standardized corporate entity matching algorithms. Standardized confidence intervals. Standardized unique identifiers. Sector, country and company type norms for corporate hierarchy.
Data Completeness	Degree and edge distributions. Stratified data comparisons with known distributions. Financial variable correlations. Manual sampling and checking of edge completeness.	Sector, country and company type norms for node, edge and financial variable distributions.
Data Accuracy	Degree and edge distributions. Outlier frequency.	Sector, country and company type norms for node and edge distributions.

To be clear, we are not claiming that work that does not give an indisputable answer to the proposed set of questions should remain unpublished. However, like caution exercised when drawing conclusions from correlations with large variance, we call on researchers to exhibit significant awareness when interpreting corporate network analysis results when these data are subject to issues around data completeness and accuracy.

A better knowledge of standard distributions of nodes, edges, financial variables and typical network structures in various corporate settings would greatly enhance assessment of the validity of BCND analyses in the future. Because of the differing competitive and regulatory imperatives in these settings, these vary by sector, company type and country. Because of the profound impact of entity ambiguity on the shape of corporate networks, and consequent validity of any analysis, there is also a need for greater standardization of entity disambiguation methods. There will be great benefits from the development of standardized corporate entity matching algorithms with explicit confidence intervals. Further progress in the development of

databases with unique identifiers and positions in corporate hierarchy will also aid this process.

Ultimately, the goal of a big data approach is to extract value. For us, this translates to knowledge. Whether a big data approach to questions related to corporate networks provides additional insight compared with studying small data remains a key question. Indeed, even if one effectively avoids all the potential problems with BCND discussed here, lacking a compelling justification for undertaking the analysis is still problematic. That the availability of big data means that one *can* conduct a given analysis is insufficient reason that one *should* conduct it. Ultimately, the proof of the pudding is in the eating, and with this intervention we hope to contribute to a vivid, candid and critical academic debate on the merits and pitfalls of using big corporate network data. In our view, early studies utilizing BCND have already led to revealing insights, for instance into the elevated level of concentration of global corporate control (Vitali et al. 2011); into the hitherto disregarded multilevel nature of board interlock networks (Heemskerk et al. 2016), and into the unprecedented shareholder power position of the Big Three passive investors in global equity markets (Fichtner et al. 2017). The promise of big corporate network data, however, goes well beyond these arguably rather descriptive contributions. Crucial next steps include understanding the driving forces behind network dynamics by utilizing advanced modelling frameworks for big data, and ultimately pinpointing the economic, political and societal consequences of the newly uncovered patterns. One cannot make these contributions systematically without first addressing key challenges associated with BCND problems.

Acknowledgements

This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant no. 638946), and also from the Russell Sage Foundation (grant no. 83-15-13).

Note

1. We are grateful to an anonymous referee for suggesting this example.

References

Agrawal, D., P. Bernstein, E. Bertino, S. Davidson, et al. (2014) 'Challenges and opportunities with big data, a community white paper', mimeo, available at: http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf.

Allen, M. P. (1974) 'The structure of interorganizational elite cooptation: interlocking corporate directorates', *American Sociological Review*, 393–406, available at: www.jstor.org/stable/2094297.

Barabási A. and R. Albert (1999) 'Emergence of scaling in random networks', *Science*, 286 (5439), 509–12, doi: 10.1126/science.286.5439.509.

Batagelj, V. and M. Zaveršnik (2003) 'An *O(m)* algorithm for cores decomposition of networks', *arXiv*, available at: https://arxiv.org/abs/cs/0310049.

Battiston, S., J. D. Farmer, A. Flache, D. Garlaschelli, A. G. Haldane, H. Heesterbeek, C. Hommes, C. Jaeger, R. May and M. Scheffer (2016) 'Complexity theory and financial regulation', *Science*, 351 (6275), 818–19, doi: 10.1126/science.aad0299.

Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar and S. Fienberg (2003) 'Adaptive name matching in information integration', *Ieee Intelligent Systems*, 18 (5), 16–23, doi: 10.1109/MIS.2003.1234765.

Blackwell, M., J. Honaker and G. King. (2015a) 'A unified approach to measurement error and missing data: details and extensions', *Sociological Methods and Research*, online article, doi: 10.1177/0049124115589052.

Blackwell, M., J. Honaker and G. King. (2015b) 'A unified approach to measurement error and missing data: overview and applications', *Sociological Methods and Research*, online article, doi: 10.1177/0049124115585360.

Borgatti, S. P., K. Carley and D. Krackhardt (2006) 'On the robustness of centrality measures under conditions of imperfect data', *Social Networks*, 28 (2), 124–36, doi: 10.1016/j.socnet.2005.05.001.

Borgatti, S. P., A. Mehra, D. Brass and G. Labianca (2009) 'Network analysis in the social sciences', *Science*, 323 (5916), 892–5, doi: 10.1126/science.1165821.

Burris, V. (2005) 'Interlocking directorates and political cohesion among corporate elites', *American Journal of Sociology*, 111 (1), 249–83, doi: 10.1086/428817.

Butts, C. (2009) 'Revisiting the foundations of network analysis', *Science*, 325 (5939), 414–16, doi: 10.1126/science.1171022.

Carpenter, M. A., M. Li and H. Jiang (2012) 'Social network research in organizational contexts a systematic review of methodological issues and choices', *Journal of Management*, 38 (4), 1328–61, doi: 10.1177/0149206312440119.

Carroll, W. K. and M. Fennema (2004) 'Problems in the study of the transnational business community', *International Sociology*, 19 (3), 369–78, doi: 10.1177/0268580904045346.

Carroll, W. K. and J. P. Sapinski (2010) 'The global corporate elite and the transnational policy-planning network, 1996–2006', *International Sociology*, 25 (4), 501–38, doi: 10.1177/0268580909351326.

Carroll, W. K., M. Fennema and E. M. Heemskerk (2010) 'Constituting corporate Europe: a study of elite social organization', *Antipode*, 42 (4), 811–43, doi: 10.1111/j.1467-8330.2010.00777.x.

Chu, J. and G. Davis (2016) 'Who killed the inner circle? The decline of the American corporate interlock network', *American Journal of Sociology*, 122 (3), 714–54, doi: 10.1086/688650.

Cohen, W. W., P. Ravikumar and S. E. Fienberg (2003) 'A comparison of string distance metrics for name-matching tasks', *Proceedings of IJCAI-03 workshop on information integration*, August, available at: www.cs.cmu.edu/~wcohen/postscript/ijcai-ws-2003.pdf.

Compston, H. (2013) 'The network of global corporate control: implications for public policy', *Business and Politics*, 15 (3), 357–79, doi: 10.1515/bap-2012-0049.

Costenbader, E. and T. W. Valente (2003) 'The stability of centrality measures when networks are sampled', *Social Networks*, 25 (4), 283–307, doi: 10.1016/S0378-8733(03)00012-1.

Cranmer, S. J. and B. A. Desmarais (2011) 'Inferential network analysis with exponential random graph models', *Political Analysis*, 19 (1), 66–86, doi: 10.1093/pan/mpq037.

Cranmer, S. J. and J. M. Gill (2013) 'We have to be discrete about this: a non-parametric imputation technique for missing categorical data', *British Journal of Political Science*, 43 (2), 425–49, doi: 10.1017/S0007123412000312.

Cronin, B. (2011) 'Networks of corporate power revisited', *Procedia: Social and Behavioral Sciences*, 10 (5), 43–51, doi: 10.1016/j.sbspro.2011.01.007.

Cronin, B. (2012) 'Transnational and national structuring of the British corporate elite', in G. Murray and J. Scott (eds) *Financial elites and transnational business: who rules the world?*, Cheltenham: Edward Elgar, 177–92.

Davis, G. F. (1991) 'Agents without principles? The spread of the poison pill through the intercorporate network', *Administrative Science Quarterly*, 36 (4), 583–613, doi: 10.2307/2393275.

Davis, G. F., M. Yoo and W. E. Baker (2003) 'The small world of the American corporate elite, 1982–2001', *Strategic Organization*, 1 (3), 301–26, doi: 10.1177/14761270030013002.

Desmarais, B. A. and S. J. Cranmer (2012) 'Statistical inference for valued-edge networks: the generalized exponential random graph model', *PLOS ONE*, 7 (1), doi: 10.1371/journal.pone.0030136.

Fan, W. and A. Bifet (2012) 'Mining big data: current status, and forecast to the future', *SIGKDD Explorations*, 14 (2), available at: www.kdd.org/exploration_files/V14-02-01-Fan.pdf.

Fichtner, J., E. M. Heemskerk and J. Garcia-Bernardo (2017) 'Hidden power of the Big Three? Passive index funds, re-concentration of corporate ownership, and new financial risk', *Business and Politics*, 19 (2), 298–326, doi: 10.1017/bap.2017.6.

Fisher, D., R. DeLine, M. Czerwinski and S. Drucker (2012) 'Interactions with big data analytics', *Interactions*, 19 (3), 50–9, doi: 10.1145/2168931.2168943.

Garcia-Bernardo, J. and F. W. Takes (2016) 'The effects of data quality on the analysis of corporate board interlock networks', *arXiv*, available at: https://arxiv.org/pdf/1612.01510.pdf.

Garcia-Bernardo, J., J. Fichtner, F. W. Takes and E. M. Heemskerk (2017) 'Uncovering offshore financial centers: conduits and sinks in the global corporate ownership network', *Scientific Reports*, 7, article number 6246, doi: 10.1038/s41598-017-06322-9.

Glavic, B. (2014) 'Big data provenance: challenges and implications for benchmarking', in T. Rabl, M. Poess, C. Baru and H.-A. Jacobsen (eds) *Specifying big data benchmarks*, Heidelberg: Springer, 72–80.

Haberly, D. and D. Wojcik (2015) 'Earth incorporated: centralization and variegation in the global company network', available at: https://ssrn.com/abstract=2699326.

Handcock, M. S. and K. J. Gile (2010) 'Modeling social networks from sampled data', *Annals of Applied Statistics*, 4 (1), 5–25, doi:10.1214/08-AOAS221.

Haunschild, P. R. (1993) 'Interorganizational imitation: the impact of interlocks on corporate acquisition activity', *Administrative Science Quarterly*, 38 (4), 564–92, doi: 10.2307/2667030.

Heemskerk, E. M and F. W. Takes (2016) 'The corporate elite community structure of global capitalism', *New Political Economy*, 21 (1), 90–118, doi: 10.1080/13563467.2015.1041483.

Heemskerk, E. M., G. Daolio and M. Tomassini (2013) 'The community structure of the European network of interlocking directorates 2005–2010', *PloS One*, 8 (7), doi: 10.1371/journal.pone.0068581.

Heemskerk, E. M., F. W. Takes, J. Garcia-Bernardo and M. J. Huijzer (2016) 'Where is the global corporate elite? A large-scale network study of local and nonlocal interlocking directorates', *Sociologica*, 2, 1–31, doi: 10.2383/85292.

Henriksen, L. F., C. H. Ellersgaard and A. G. Larsen (2016) 'Stability and change in corporate governance networks', paper presented at the INSNA Sunbelt conference, April.

Huisman, M. (2009) 'Imputation of missing network data', *Journal of Social Structure*, 35 (4), 1–29, available at: www.cmu.edu/joss/content/articles/volume10/huisman.pdf.

Jagadish, H. V. (2015) 'Big data science: myths and reality', *Big Data Research*, 2 (2), 49–52, doi: 10.1016/j.bdr.2015.01.005.

Kim, J. W., B. Kogut and J. S. Yang (2015) 'Executive compensation, fat cats, and best athletes', *American Sociological Review*, 80 (2), 299–328, doi: 10.1177/0003122415572463.

Kentor, J. and Y. S. Jang (2004) 'Yes, there is a (growing) transnational business community: a study of global interlocking directorates 1983–98', *International Sociology*, 19 (3), 355–68, doi: 10.1177/0268580904045345.

Kentor, J. and Y. S. Jang (2006) 'Different questions, different answers: a rejoinder to Carroll and Fennema', *International Sociology*, 21 (4), 602–6, doi: 10.1177/0268580906065303.

Koskinen J. H., G. L. Robins and P. E. Pattison (2010) 'Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation', *Statistical Methodology*, 7, 366–84, doi: 10.1016/j.stamet.2009.09.007.

Koskinen J. H., G. L. Robins, P. Wang and P. E. Pattison (2013) 'Bayesian analysis for partially observed network data, missing ties, attributes and actors', *Social Networks*, 35 (4), 514–27, doi: 10.1016/j.socnet.2013.07.003.

Kossinets, G. (2006) 'Effects of missing data in social networks', *Social Networks*, 28 (3), 247–68, available at: https://arxiv.org/pdf/cond-mat/0306335.pdf.

Laney, D. (2001) '3D management: controlling data volume, velocity, and variety', technical report 949, META Group, available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Laumann, E. O., P. V. Marsden and D. Prensky (1989) 'The boundary specification problem in network analysis', in L. C. Freeman, D. R. White and A. K. Romney (eds) *Research methods in social network analysis*, Fairfax, VA: George Mason University Press, 61–87.

Leifeld, P., S. J. Cranmer and B. A. Desmarais (2017) 'Temporal exponential random graph models with btergm: estimation and bootstrap confidence intervals', *Journal of Statistical Software*, forthcoming.

Little, R. and D. Rubin (1987) *Statistical analysis with missing data*, Hoboken: John Wiley & Sons.

Marple, T., B. Desmarais and K. Young (2017) 'Collapsing corporate confusion: leveraging network structures for effective entity resolution in corporate relational data', available at: https://ssrn.com/abstract=3053632.

Marsden, P. V. (1990) 'Network data and measurement', *Annual Review of Sociology*, 16, 435–63, doi: 10.1146/annurev.so.16.080190.002251.

Mayer-Schönberger, V. and K. Cukier (2013) *Big data: a revolution that will transform how we live, work, and think*, Boston: Mariner Books.

Mestres, J., E. Gregori-Puijané, S. Valverde and R. Solé (2008) 'Data completeness: the Achilles heel of drug-target networks', *Nature Biotechnology*, 26 (9), 983–4, doi: 10.1038/nbt0908-983.

Mizruchi, M. S. (1989) 'Similarity of political behavior among large American corporations', *American Journal of Sociology*, 95 (2), 401–24, doi: 10.1086/229274.

Mizruchi, M.S. (1996) 'What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates', *Annual Review of Sociology*, 22 (1), 271–98, doi: 10.1146/annurev.soc.22.1.271.

Murray, J. (2014) 'Evidence of a transnational capitalist class-for-itself: the determinants of PAC activity among foreign firms in the Global Fortune 500, 2000–2006', *Global Networks*, 14 (2), 230–50, doi: 10.1111/glob.12037.

Neal, Z. (2014) 'The backbone of bipartite projections: inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors', *Social Networks*, 39 (1), 84–97, doi: 10.1016/j.socnet.2014.06.001.

Ouzienko, V. and Z. Obradovic (2013) 'Imputation of missing links and attributes in longitudinal social surveys', *Machine Learning*, 95 (3), 329–56, doi: 10.1007/s10994-013-5420-1.

Rao, H. and K. Sivakumar (1999) 'Institutional sources of boundary-spanning structures: the establishment of investor relations departments in the Fortune 500 industrials', *Organization Science*, 10 (1), 27–42, doi: 10.1287/orsc.10.1.27.

Rubin, D. (1976) 'Inference and missing data', *Biometrika*, 63 (3), 581–92, doi: 10.1093/biomet/63.3.581.

Rubin, D. (1987) *Multiple imputation for nonresponse in surveys*, New York: Wiley.

Salton, G., A. Wong and C. S. Yang (1975) 'A vector space model for automatic indexing', *Communication of the ACM*, 18 (11), 516–620, doi: 10.1145/361219.361220.

Schweitzer, F., G. Fagiolo, D. Sornette, F. Vega-Redondo, A. Vespignani and D. R. White (2009) 'Economic networks: the new challenges', *Science*, 325 (5939), 422–5, doi: 10.1126/science.1173644.

Squartini, T., I. van Lelyveld and D. Garlaschelli (2013) 'Early-warning signals of topological collapse in interbank networks', *Scientific Reports*, 3 (3357), 1–9, doi: 10.1038/srep03357.

Starrs, S. (2013) 'American power hasn't declined: it globalized! Summoning the data and taking globalization seriously', *International Studies Quarterly*, 57 (4), 817–30, doi: 10.1111/isqu.12053.

Stokman, F. N., R. Ziegler and J. Scott (1985) *Networks of corporate power*, Cambridge: Polity Press.

Takes, F.W. and E. M Heemskerk (2016) 'Centrality in the global network of corporate control', *Social Network Analysis and Mining*, 6 (97), 1–18, doi: 10.1007/s13278-016-0402-5.

Tuschke, A., W. G. Sanders and E. Hernandez (2014) 'Whose experience matters in the boardroom? The effects of experiential and vicarious learning on emerging market entry', *Strategic Management Journal*, 35 (3), 398–418, doi: 10.1002/smj.2100.

Useem, M. (1984) *The inner circle: large corporations and the rise of business political activity in the US and UK*, New York: Oxford University Press.

van der Pijl, K., O. Holman and O. Raviv (2011) 'The resurgence of German capital in Europe: EU integration and the restructuring of Atlantic networks of interlocking directorates after 1991', *Review of International Political Economy*, 18 (3), 384–408, doi: 10.1080/09692290.2010.488454.

Vitali, S., J. B. Glattfelder and S. Battiston (2011) 'The network of global corporate control', *PloS One*, 6 (10), 1–6, doi: 10.1371/journal.pone.0025995.

Wang C., C. T Butts, J. R. Hipp, R. Jose, C. M. Lakon (2016) 'Multiple imputation for missing edge data: a predictive evaluation method with application to add health', *Social Networks*, 45, 89–98, doi: 10.1016/j.socnet.2015.12.003.

Ward, M. D., P. D. Hoff and C. L. Lofdahl (2003) 'Identifying international networks: latent spaces and imputation', in R. Breiger, K. Carley and P. Pattison (eds) *Dynamic Social network modeling and analysis: workshop summary and papers*, Washington, DC: National Academies Press, 345–59.

Winkler, W. E. (1990) 'String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage', *Proceedings of the section on survey research methods (American Statistical Association)*, 354–59, available at: https://goo.gl/Pdhv5S.

Young, K., T. Marple and J. Heilman (2017) 'Beyond the revolving door: advocacy behavior and social distance to financial regulators', *Business and Politics*, 19 (2), 327–64, doi: 10.1017/bap.2017.10.

Žnidaršič, A., A. Ferligoj and P. Doreian (2012) 'Non-response in social networks: the impact of different non-response treatments on the stability of blockmodels', *Social Networks*, 34.