Information Systems 000 (2017) 1–9

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

The effects of data quality on the analysis of corporate board interlock networks

Javier Garcia-Bernardo[a,*], Frank W. Takes[a,b]

[a] CORPNET, University of Amsterdam, Amsterdam, The Netherlands
[b] LIACS, Leiden University, Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 15 April 2017
Revised 12 September 2017
Accepted 15 October 2017
Available online xxx

ABSTRACT

Nowadays, social network data of ever increasing size is gathered, stored and analyzed by researchers from a range of disciplines. This data is often automatically gathered from API's, websites or existing databases. As a result, the quality of this data is typically not manually validated, and the resulting social networks may be based on false, biased or incomplete data. In this paper, we investigate the effect of data quality issues on the analysis of large networks. We focus on the global board interlock network, in which nodes represent firms across the globe, and edges model social ties between firms – shared board members holding a position at both firms. First, we demonstrate how we can automatically assess the completeness of a large dataset of 160 million firms, in which data is missing not at random. Second, we present a novel method to increase the accuracy of the entries in our data. By comparing the expected and empirical characteristics of the resulting network topology, we develop a technique that automatically prunes and merges duplicate nodes and edges. Third, we use a case study of the board interlock network of Sweden to show how poor quality data results in distorted network topologies, incorrect community division, biased centrality values and abnormal influence spread under a well-known diffusion model. Finally, we demonstrate how the proposed data quality assessment methods help restore the network structure, ultimately allowing us to derive meaningful and correct results from the analysis of the network.

1. Introduction

Over the past few decades, the amount of digital information has been doubling roughly every two years. At the same time, there is an ongoing and prevailing desire to extract meaningful knowledge from this data. Although many knowledge discovery methods and techniques are scalable to larger volumes of data, “big data” [1] has the significant and largely unaddressed problem of “veracity” [2]. This refers to the fact that the explosion in the amount of available data has resulted in a situation in which researchers can no longer manually validate the *quality* of their data [3]. Data quality most dominantly relates to questions of *completeness* (what part of the data do we have, and what part do we miss?) and *accuracy* (is the data that we have correct and suitable for answering our particular domain questions?). Issues with data quality have an estimated cost ranging from $611 billion [4] to $3.1 trillion (IBM estimate) per year in the United States alone. As such, increased importance has been given to data quality in the information systems literature [5]. Assessing and correcting data quality issues (data cleaning) is a main pre-processing step often required in the analysis data [6]. Here we specifically set out to assess how these issues can be addressed in the context of (social) network analysis.

In this paper we focus on so-called *corporate networks*, in which links represent particular relationships between corporations. Ties in corporate networks can be based on a variety of relationships between firms, including trade [7], borrowing and lending of money [8], ownership [9], or as we analyze in this paper: shared board members. In these networks of *interlocking directorates*, also referred to as *board interlock networks*, a node represents a firm and an edge between two firms denotes that these firms share at least one board member or director. An example of a board interlock network is given in Fig. 1. Board interlocks are common practice in today's corporate world, and over the past century, social scientists have extensively studied the causes and consequences of board interlocks. See for example the excellent overview given in [10], where Mizruchi discusses how interlocks relate to collusion, monitoring (e.g., banks keeping an eye on firms they invested in), legitimacy (attracting board members with a particular reputa-

[*] Corresponding author.
E-mail addresses: garcia@uva.nl (J. Garcia-Bernardo), takes@uva.nl (F.W. Takes).

https://doi.org/10.1016/j.is.2017.10.005
0306-4379/© 2017 Elsevier Ltd. All rights reserved.

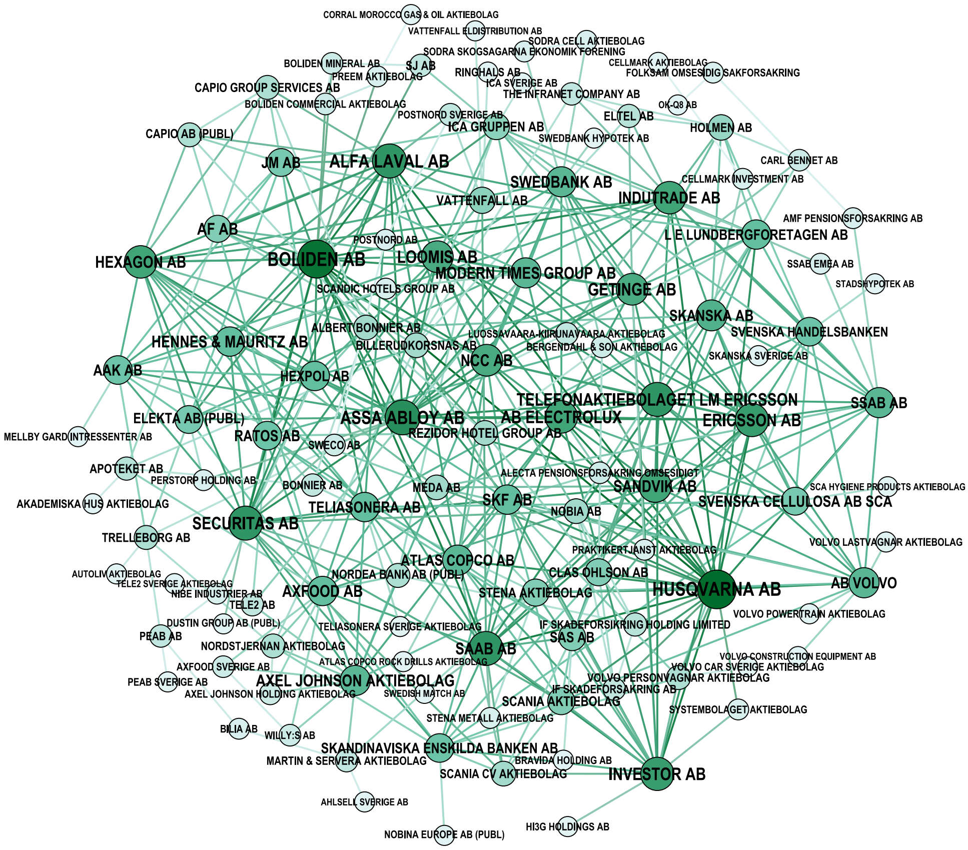

Fig. 1. Sample of the Swedish network of interlocking directorates. Based on largest 120 firms in terms of revenue, connected through 422 interlocks. Darker color and larger size denote a higher node degree. Layout created in Gephi (http://gephi.org) using the ForceAtlas2 algorithm.

tion in a certain area that is of importance to the firm), individual career advancement and social cohesion (social ties among the upper class). Previous research has established that networks of interlocking directorates facilitate the spread of governance routines and practices, the exchange of resources, communication and the dissemination of new ideas [11,12]. Since a significant number of directors has positions at two or more firms, the board meetings of these firms connect the majority of big businesses in the world. For instance, Davis [12] discusses how the majority of the corporate elite would rapidly be infected by a contagious disease as a result of the small world property of the network of interlocking directorates.

Corporate networks have interesting topological characteristics common to all real-world networks, including a fat tailed degree distribution, the emergence of a giant component and very low pairwise distances between nodes [13]. Researchers have thus applied established social network analysis methods and techniques [14] to corporate networks. Community detection has been used to understand the geographical dimension of the structure of these networks [15,16] and centrality measures provide insight into powerful firms and countries in the network [17].

Initially, social scientists studied only small networks of interlocking directorates, typically based on a few hundred firms and their relationships. Researchers carefully double-checked the data they manually gathered from annual reports of the companies involved. Nowadays, large databases on corporations are provided by commercial corporate information providers such as Orbis, BoardEx, ThomsonOne and Bloomberg, including information on their financial performance and board composition. The availability of these databases with millions of firms allows our board interlock networks to be automatically constructed from the available firm and board member data. Here, we focus on a dataset extracted from Orbis, a large information provider that gathers data from country registers across the world, and then makes this data available through one database (for details, see Section 2). The sheer volume of contemporary corporate network data [18] means that it is no longer possible to manually check each firm, let alone their board composition, for correctness. However, the quality of our data is diverse across countries and country registers. Indeed, two problems of data quality (completeness and accuracy) come into play here. Note that other common aspects of data quality [19], such as illegal values and varying value representations, are already corrected by the information provider.

The first data quality issue, *completeness*, is not necessarily problematic, for example when we have a dataset in which data is missing completely at random (MCAR), or when missing values are directly correlated with a known variable (MAR). We find that in our dataset information about some attributes (e.g., number of employees) is correlated with other attributes with better availability – it is MAR. However, often it is not the attributes but the companies themselves that are not present in the data, meaning that data is missing *not* at random (MNAR). This may result in severe problems, because if non-random parts of the data are missing, we can no longer consider it a reliable sample and derive meaningful results for the system represented by the dataset as a whole. For example, if we blindly use the data provided by information provider Orbis to compare countries, we observe that the average Mexican firm is larger than the average firm in the United States. In turns out that this is due to lower data quality in Mexico, where many small companies are not included in the data, thus increas-