Universiteit Leiden
Opleiding Informatica

Analyzing the resilience of
board interlock networks
under imperfect data

Name:	Roberto Lucchese
Date:	14/10/2016
1st supervisor:	Frank W. Takes
2nd supervisor:	Walter A. Kosters
3rd supervisor:	Eelke M. Heemskerk (UvA)

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Acknowledgments

I would really like to thank all the CORPNET group at the University of Amsterdam (see `http://corpnet.uva.nl`) for guiding me towards a better understanding of the subject, the data and the methods. In particular, a special thanks goes to Frank W. Takes, in the role of my first supervisor, for his guidance throughout the work, for his patience and his insightful comments.

Abstract

Network science is gaining increasingly more attention. It allows us to transform large quantities of flat data into easier and more understandable models of objects (or *nodes*) and relationships between these objects (or *edges*). Datasets, though, usually suffer from data quality issues. Problems with "completeness" and "accuracy" are just two of the many issues that are usually present.

Here we study the effect of these data quality issues on the analysis of *corporate board interlock networks*, in which nodes represent companies and the edges are shared board members.

More accurate and complete information on larger companies, misspelled names, spurious companies and connections are some of the data quality artifacts one may encounter working with corporate data. To understand their impact we stress 6 networks under 15 different data quality artifacts and we study the changes in some of the most frequently used network measures.

Our results suggest that despite imperfect data quality, the observed networks remain very similar to the original ones under most of the artifacts. We show how community analysis is barely influenced by most data quality issues and how degree centrality is far more resistant than betweenness and harmonic centrality. Finally, we conclude that board interlock networks are resilient enough to still be studied under most data quality issues.

Contents

1 Introduction

The amount of data available to researchers and industries has grown enormously. Given this overabundance, manually inspecting and understanding data has become a difficult task. The need for a simple and effective model has pushed researchers and industries towards network science. The advantage in using networks as a model is the ability to transpose very complex systems and problems into understandable and easy to analyze models of objects (nodes) and relationships between these objects (called edges). For this reason network science is very much under the attention of many fields of studies such as mathematics, physics, computer science and computational social science. Where the first two concentrate their attention on understanding the mathematical properties, the latter two are usually more attracted to the combination of network science and big data analytics: modeling large data sources as networks and deriving empirical conclusions analyzing their features (see Figure 1).

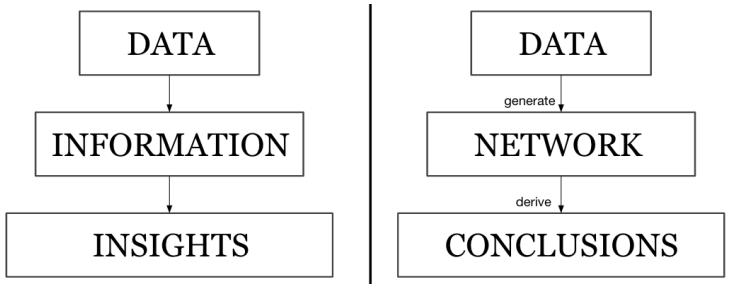


Figure 1: From data to insights using networks.

Big datasets, though, usually present numerous kinds data quality issues, such as problems with "consistency", "relevancy", "interpretability", "completeness" and "accuracy". Many others are presented in Pipino et al. (2002). If we translate these large sources of data into networks, how do these data quality issues effect their derived analysis?

Here we want to answer this question concentrating on corporate board interlock networks, in which the nodes represent companies and the edges are based on the board members these companies share. The topological and structural features of the network representation of such data gives new insights in the economical and political aspects of the system. But, are they resilient enough to still be studied when the data quality is poor?

Retrieving corporate information usually goes through local and global providers. Fewer global providers retrieve information. Misspelled companies or spurious connections are indeed often found when working with corporate data. Furthermore, where other works find a way to identify and add missing nodes or links (Kim and Leskovec (2011)), our data is too complicated and rich of information to simply impute it. Attributes such as number of employees, revenue and geographical position of the companies are just a few of the numerous attributes that our data has. This work, thus, will be focused on understanding at which level imperfect data sources really affect the final analysis of board interlock networks.

We start from the work of Borgatti et al. (2006) and in particular of Wang et al. (2012) where six definitions of data quality artifacts are given, namely: node removal, edge removal, node addition, edge addition, node aggregation and node disaggregation. We apply each of these scenarios to six corporate board interlock networks with the goal to understand how each scenario influences them. Where all the aforementioned scenarios worked by randomly selecting nodes or edges, here we will introduce biases both on the degree and on the revenue of the companies. Most of the times, indeed, the artifacts present in datasets do not occur uniformly at random, but instead, they are biased towards some property of the data. More accurate and complete data on larger firms, for instance, is just one of the problems we may encounter working with corporate datasets.

We proceed by analyzing the changes in many of the most frequently used network measures in corporate network analysis: degree distribution, distance distribution, degree centrality, betweenness centrality, harmonic (closeness) centrality, density, average distance, assortativity and clustering coefficient. Eventually, we study the changes in community partitions by means of the variation of information (Meilă, 2007).

We derive empirical conclusions regarding the resilience of the Italian, Danish, UK, Scandinavian, Spanish and Dutch corporate board interlock network's giant components, characterizing each of them using a simple and effective matrix: the *resilience matrix*. We define this *resilience matrix* as a matrix describing how "different" a perturbed network is from its original version, when errors are introduced. The concept of "difference" is given by the changes in the measures listed above. The lower are the changes, the higher is the resilience. Ideally, a completely resilient network will have the same features both before and after the artifacts are applied.

We conclude this work reasoning about whether the resilience of the net-

works is sufficient to guarantee significant studies even under poor data quality conditions.

This thesis is organized as follows: In Section 2 we present graph theory and we discuss some synthetic models as generators of real-world networks. In Section 3 we present related work, both regarding the analysis of corporate board interlock networks and on the effect of data quality in social networks. In Section 4 we present the data, while in Section 5 we discuss the artifacts and the measurements of the error we use in this work. In Section 6 we present the results of our analysis. In a detailed way first and then by means of the resilience matrix. In Section 7 we draw conclusions.

2 Network Science

One of the most common and widespread models to shape interactions between objects in data are so-called complex networks. These networks can be seen as graphs in which the nodes represent objects and the edges represent some kind of interaction between these objects. Modeling small or large-scale data as networks allow us to examine how objects act and interact with others, to study *direct* and *indirect* interactions, to compute communicational potentials and to detect communities and the most important actors throughout the system. In the subsections that follows we present networks (or graphs) in their theoretical aspects. We start by introducing graph theory and presenting some of the most used network measures. We then proceed by presenting real-world network properties and we finally discuss whether synthetic models are feasible as generators of real-world graphs.

2.1 Graph Theory

Let us call $G = (V^G, E^G) = (V, E)$ a *graph* (or *network*) where V is the set of *vertices* (or *nodes*) and E the set of *edges* (or *links*) connecting pairs of vertices. We define $n = |V|$ as the number of nodes and $m = |E|$ as the number of edges. Here we always consider an *undirected* network where if there exists a link (i, j) from node i to node j, there always exists a link (j, i) from node j to node i too. We also consider a network without *self-loops*. No edges can start and end in the same node. Associating then with every edge $\{i, j\} \in E$ or (i, j) a value $w \in \mathbb{Z}^+$ called the *weight*, we define a *weighted* network. Given a node i, we define its *neighbors* $N(i)$ as set of nodes incident to i and $ddeg(i) = |N(i)|$ as its *degree*. If we also sum the weights of the edges going from i to all its neighbors, we are defining the *strength* $s(i)$ of the node i. We then call a *path* between any two vertices i and j, a sequence of edges connecting i and j and we define the *distance* $d(i, j)$ as the number of steps in a shortest path between i and j. If we compute the average number of steps in the shortest paths that connect all possible pairs of nodes, we are defining the *average distance*. Eventually, we call the *degree distribution* $P(k)$ of a graph G, the distribution of the fraction of nodes with degree k. If we consider this distribution in a logarithmic scale, we often obtain approximately a straight line. The slope s of this latter line, in real-world networks, is usually in the range $2 < s < 3$. The *distance distribution* $D(z)$, instead, is defined as the fraction of pairs of nodes at distance z.

Let us now define a *subgraph* S of a graph G, as a graph whose vertices and edges are a subset of the vertices and edges of G, and where only edges connecting nodes in the subset are present. Given an *undirected* graph G, we then define the *connected components* of G as the maximal subgraphs in which every node is connected to at least another node of the same subgraph and there is no edge connecting nodes between subgraphs. The connected component that contains the majority of the nodes is defined as the *giant component* S_G of the graph.

In the field of social sciences, one of the most studied properties of a network is the *global clustering coefficient*, also called *transitivity* in case of a directed network. It is based on the concept that if a person i is friend of a person j and a person j is friend of a person z, there is a high probability that person i and person z are also friends (Newman, 2003b). Person i, j and z will then form a "triangle". Formally, the global clustering coefficient is defined as follows:

$$C = \frac{3 \times \textit{number of triangles}}{\textit{number of connected triples}} \tag{1}$$

where with *connected triplet* we refer to three nodes connected by two edges.

Let us now take one of the most used social networks of these days, Facebook. It counts millions of users but each one has a relatively low number of friends compared to the total number of people that (actively or not) use it. More formally, if we measure the ratio between the actual connections and the potential connections of a network, we are measuring the *density* of the graph. In formula:

$$d = \frac{m}{n(n-1)/2} \tag{2}$$

Very large real-world networks such as Facebook, for instance, typically have really low densities.

Finally, considering Facebook again, one might be interested in computing the *scalar assortativity coefficient*: a value $r \in [-1; 1]$ measuring how much people with similar degree interact with each other. More formally, the scalar assortativity coefficient (Newman, 2003a) is defined as the Pearson correlation coefficient between the degrees of pairs of connected nodes. It simply measures the linear correlation between types of degrees. Eventually, $r = -1$ implies complete disassortativity, $r = 1$ complete assortativity and $r = 0$ non-assortativity.

2.2 Centrality Measures

Measures of centrality allow us to find the most central actors in the system. Naturally, "centrality" can have numerous meanings. Here we present four of the most used measures of centrality:

- Degree centrality: presented by Freeman (1979), it defines the "communication activity" of each node. Given a node i, its degree centrality value will be computed as follows:

$$C_d(i) = \frac{deg(i)}{n-1} \tag{3}$$

  What distinguishes Equation 3 from the basic definition of degree is that here we have a normalized measure in the range $[0, 1]$.

- Betweenness centrality: also elaborated by Freeman (1977), it is an important indicator of the nodes that act as bridges. In particular, where high degree nodes are "important" because the high number of connections, high betweenness centrality nodes are "important" because of their strategic position: they lie in *between* other actors. For this reason these strategic nodes are also known as the "brokers". More generally, *betweenness centrality* is defined as follows:

$$C_b(i) = \sum_{j \neq y \neq i \neq j} \frac{\sigma_{jy}(i)}{\sigma_{jy}} \text{ with } i, j, y \in V \tag{4}$$

  We call $\sigma_{jy}(i)$ the number of shortest paths from node j to node y that pass through i, while σ_{jy} is the total number of shortest paths from j to y. Dividing then $C_b(i)$ by $\frac{1}{2}(n-1)(n-2)$ we obtain a normalized measure.

- Closeness centrality: this measure highlights *central* nodes or, in other words, it highlights the nodes having lower distance from all the others (Freeman, 1979). More formally, closeness centrality defines how a node i is close to all the other nodes in the network:

$$C_c(i) = \frac{n-1}{\sum_j d(i,j)} \text{ with } i, j, y \in V \tag{5}$$

- Harmonic centrality: closeness centrality is based on the concept of distance and it defines how close a node is to all the other nodes in the network. In its original definition it is not applicable in disconnected graphs. The distance between any two disconnected nodes is infinite. In order to avoid this problem a different definition of closeness centrality, called *harmonic centrality*, was introduced by Rochat (2009). Harmonic centrality is formally defined as follows:

$$H_c(i) = \sum_j \frac{1}{d(i,j)} \text{ with } i, j, y \in V \tag{6}$$

with $d(i, j)$ the shortest-path distance between i and j in the network. Now, the distance between any two disconnected nodes will contribute for $\frac{1}{\infty} = 0$, instead of infinity. So, the contribution of an unconnected node to the harmonic centrality value of a node i will now be zero and not infinite.

2.3 Community detection

As a result of the increased computational power of computers and servers, important properties of real-world networks have been revealed. Many of them are *small-world* networks, meaning that the average path length from any two nodes is surprisingly low (6.6 in the Microsoft IM studied by Leskovec and Horvitz (2007)) while their clustering coefficient is relatively high (Watts and Strogatz, 1998). Another important property is the so-called *scale-free* property: in many real-world networks, the degree distribution follows a power-law (Barabási and Albert, 1999). Finally, a clear community structure (the actors of the systems are well divided into non-overlapping groups) is often present (Fortunato (2010)).

In real life, talking about *communities* (or *clusters*) we usually refer to groups of people sharing common interests or, more generally, groups of people that tend to interact more between themselves than with other people.

In network analysis the concept of *communities* is similar, namely: in the system, small or large groups of objects that are more densely connected internally than externally, are often present. However, the problem of finding communities, of any size, in real-world graphs, is not an easy task. Here we present the community detection problem as an optimization problem and we will present one of the fastest and most used algorithm able to unveil the community structure of large and very large networks in $O(n \log n)$ time.

2.3.1 Modularity & Louvain method

Looking for a measure able to quantify the quality of their community detection algorithm, Newman and Girvan (Newman and Girvan, 2004) presented *modularity*. This measure is based on the idea that only having few edges connecting communities is not enough to define a good community structure (Newman, 2006). Indeed, the role of modularity is to understand whether the number of edges connecting communities (*extra-community* links) is fewer than expected, or, in the same way, whether the number of *intra-community* links is more than expected. In other words, modularity is able to quantify the difference between the number of intra-community edges in a network and the expected number of edges withing groups in a randomly rewired network.

Let us take a weighted network W with n nodes and adjacency matrix A where each element A_{ij} represents the weight of the edge $\{i, j\}$. We then define $m = \frac{1}{2}\sum_{ij} A_{ij}$ as the sum of all the weights of the links. Given now a partition vector $\vec{C}$ of size n, we can express the modularity as follows (Blondel et al., 2008):

$$Q(\vec{C}) = \frac{1}{2m}\sum_{i,j}\Big[A_{ij} - \frac{k_i k_j}{2m}\Big]\delta(c_i, c_j) \tag{7}$$

where $\frac{k_i k_j}{2m}$ is the expected number of edges from node i to node j if they were placed at random according to the vertex strength (Chung and Lu, 2002), and $\delta(c_i, c_j)$ is the Dirac delta function which gives 1 if $c_i = c_j$ (meaning i and j are in the same community) and 0 otherwise.

The optimization of the modularity function is at the basis of many community detection methods, that, since the optimization problem is NP-complete (Brandes et al., 2008), have as goal finding good approximations. One of the most commonly used methods present in literature is the well-known Louvain method.

The Louvain method (Blondel et al., 2008) is known for finding high quality communities in large networks with very low computation time. It is divided into two different phases iteratively repeated. At the beginning of the first phase each node represents a community on its own. Then the algorithm proceeds by considering each node and computing the modularity value of the community structure obtained by placing the considered node i in the same community as j (with $j \in N(i)$). If by placing i in the

same community as one of its neighbors the modularity value (positively) increases, i is considered part of that community, otherwise it stays in its original community. Intuitively, i will be placed in the community together with the neighbor for which the modularity value is maximized. This phase ends when all nodes have been analyzed and no improvements can be obtained. In the second phase each community found in the previous phase is condensed in a single node. Intra-community edges are now represented by self-loops and any two nodes of this new network will be connected with an edge having a weight equal to the sum of the weights of the nodes that are part of those two communities. When this phase is also completed, the next iteration starts, considering the condensed network as the "starting" network. The final community structure is obtained when no more improvements, in terms of modularity value, can be made.

2.4 Real-world and Synthetic Networks

Where here, given the abundance of data, we decide to use real-world corporate data, mathematicians and theoretical network scientists are working to build always more precise synthetic examples (Van der Hofstad, 2016) that resemble all the real-world networks features. The researchers' goal is indeed to understand the fundamental characteristics a synthetic model should have in order to faithfully reproduce real-world networks features.

In order to give a more technical flavor to the interested reader, in the subsections that follow we present some of those models and we will discuss their feasibility as generators of networks that resemble real-world ones.

2.4.1 Erdős-Rényi Random Graph

One of the first and simplest examples of random graphs is the *Erdős-Rényi random graph* (*ER*) (Erdős and Rényi, 1959). The procedure is straightforward: we start with a complete network, consisting of n vertices, in which each pair of vertices is connected by an edge. We then decide to retain each edge with probability $p \in (0, 1)$, or remove it with probability $1 - p$. Despite its simple construction method, it has very interesting mathematical properties. In particular, we know that for $n \to \infty$ its degree distribution resembles a Poisson distribution (for any value of p) with the mean on the average degree. This implies that most of the nodes will have an average degree, and that is it difficult to find nodes both with really low and with

really high (greatly exceeding the average) degree. The latter nodes are more formally defined as *hubs*. Given these properties, both not true for real-world networks (given their power-law degree distribution), the *Erdős-Rényi random graph* is not suitable for constructing synthetic networks that reflect the properties of real-world networks.

2.4.2 Configuration Model

Another well-known class of random graph models is the *configuration model*. The main idea behind it, is to give as parameter to the model the observed degree sequence of an empirical network and then, in its simplest form, to create graphs compatible with that degree sequence (randomly connecting vertices). We can see the configuration model as a *null-model*, which generates graphs compatible with a given degree sequence, whose intent is to provide a benchmark to compare an empirical network to. In this way, if our empirical network has some properties in common with the synthetic graphs realized by the configuration null-model, one can conclude that those properties are only due to the degree distribution. Whether we notice different properties with respect to the null-model, possible interesting results and conclusions are worth to be investigated. In other words, if the density of the empirical network, for instance, is way different from the density of the null-model generated starting from the observed degree sequence, then a particular trend (or interesting phenomena to study) in the network might be present.

Here we look in detail at one of the most commonly used configuration models: the *Chung-Lu model* (Chung and Lu, 2002). It generates a *canonical* ensemble of graphs (all possible realizations of a network model) meaning that the constraints on the degree sequence (in this particular case) are *soft*. This implies that not all the graphs generated will have the exact same degree sequence of our empirical network, but, that the average of all the realizations of the networks will. With this important property we can, in most cases, reduce biases and simplify the study of these ensembles. Finally, all the ensembles in which the constraints are *hard* (not imposed on the average but on each single realization) are called *microcanonical* ensembles.

Given the expected degree sequence of an empirical network having m edges, the Chung-Lu model will draw an edge from node i to node j with

the following probability:

$$q_{ij} = \frac{k_i k_j}{2m} \tag{8}$$

where k_i is the degree of node i and k_j the degree of node j of our empirical network. This model generates all possible graphs (with n nodes) with a certain probability, the distribution of which is specified by all the q_{ij}. Naturally, the graphs having a completely different degree sequence, from the empirical one, will be very unlikely. Locally, connections between nodes of high degree are highly probable, and connections between nodes with low degree are less probable.

Two important remarks are necessary when using the Chung-Lu model. The first remark is about self-loops. For simplicity let us start by putting all q_{ij} in a matrix that we define as Q, in which $Q_{ij} = q_{ij}$. Now, when generating an ensemble, we usually want to obtain a collection of undirected graphs that will have no self-loops (as most of the real-world networks do). In this case, it is important to notice that if no specification on the fact that the elements in the diagonal of Q (that are normally equal to $\frac{k_i^2}{2m}$, and so different from zero) must be 0, the Chung-Lu model will produce graphs with self-loops. Despite producing graphs with self-loops does not exactly resemble the topological properties of most real-world networks, this does not influence the results (see remark in Subsection 2.4.4). Note, indeed, how even in the modularity formula (Equation 7), no such specification on avoiding self-loops is made.

The second remark is about the feasibility of the model. Let us now remember how each value of q_{ij} (being a probability) has to be in the range $[0, 1]$ for all possible i, j. If we then want to ensure this constraint, we only have to consider degree sequences in which all the nodes have degree $k_i \leq \sqrt{2m}$. Considering a node i with degree $k_i = \sqrt{2m}$, we have $q_{ii} = 1$. For this reason, $k_i = \sqrt{2m}$ is the last upper-bound in order to still have all the $q_{ij} \leq 1$. Unfortunately, this constraint violates most of the power-law degree distributions of real-world networks.

2.4.3 Preferential Attachment

Where the configuration model and the *ER* random graph model do no take explicit hypotheses on how the network organizes itself, the *preferential attachment* model starts from the idea that most real-world networks grow over time (e.g., Facebook or the WWW) and popular (high degree) users or web pages will likely become even more popular. This model, also called

Barabási-Albert model (Barabási and Albert, 1999), is known for its *"the rich get richer"* behavior. In particular, knowing that real-world networks are usually scale-free (have a power-law degree distribution), the Barabási-Albert model is able to create scale-free networks following the convention for which the higher the degree of a node is, the higher will be the probability that new nodes inserted in the network will be attached to it.

Despite that this model is able to generate scale-free networks with small average path length, it fails to reproduce the relatively high clustering coefficient usually present in real-world networks (Zafarani et al., 2014).

2.4.4 Stochastic Blockmodels

Another well-known random graph model is the Stochastic Blockmodel (Holland et al., 1983). In its simplest form, given a certain number of nodes n and a function that assigns to each vertex a value of membership in one of the possible K different communities (or blocks), this model places edges between vertices with probabilities given by the described membership assignment. In particular, vertices inside the same community will be connected with higher probability than two vertices belonging to two different communities.

One of the main applications of stochastic blockmodels is community detection (*a posteriori* block-modeling). If we define g as a vector of size n in which g_i denotes the group to which vertex i belongs and ω_{rs} as the expected value of the adjacency matrix A_{ij} of an undirected multi-graph in which vertex i belongs to group r and vertex j belongs to group s, we can maximize the probability $P(G|\omega, g)$ with respect to the unknown model parameters g and ω. This then becomes the (log-)likelihood maximization problem well described by Karrer and Newman (2011).

Unfortunately, this simplest version of the stochastic blockmodel does not generate networks with structures reflecting the ones found in real-world networks.

2.4.5 Degree-Corrected Stochastic Blockmodel

For the aforementioned reason, Karrer and Newman (2011) present a more elaborate version of the stochastic blockmodel, called the Degree-Corrected Stochastic Blockmodel. In this model the probability distribution over the networks (undirected multigraphs with self-loops) depends both on parameters g and ω, but also on a vector θ, in which θ_i contains the expected degree

of vertex i. As in its simplest version, also in the degree-corrected version of this model, the number of edges between any two nodes (multi-graph) is drawn following a Poisson distribution with mean $\lambda = \theta_i \theta_j \omega_{g_i g_j}$. From the latter statement, we notice how the expected value of A_{ij} will be equal to $\theta_i \theta_j \omega_{g_i g_j}$. The *a posteriori block-modeling* problem now, again, becomes a (log-)likelihood maximization problem, in which the $P(G|\theta, \omega, g)$ has to be maximized.

This model, taking in consideration the expected degree sequence, seems to outperform the standard stochastic blockmodel in generating networks having properties that best resemble the real-world ones. While both the Chung-Lu model and the simplest version of the Stochastic blockmodel cannot generate networks having a power-law degree distribution (despite in the latter the networks have a clear community structure), using the Degree-Corrected version, we have both the desired power-law and a clear community structure.

2.4.6 Synthetic vs Real-world networks

In this work, as previously mentioned, we decide to not use synthetic networks. The main reason for this choice is the inability of these synthetic models to reproduce the node attributes board interlock networks may have. Names, revenues, geographical locations of the companies, number of employees are all not reproducible by a synthetic model.

Eventually, synthetic models, even if important in studying how network features arise, are still imperfect in reproducing the set of features a complex real-world network has. Despite this, we strongly believe that more theoretical studies on the influence of data quality in network analysis by means of synthetic toy models, will give the mathematical and theoretical perspective on the problem that is still missing. Where empirical studies are meant to understand how data quality artifacts impact certain networks, mathematicians can contribute with theorems and proof of what must happen when an artifact is present. Starting by constructing synthetic toy models, mathematicians may then analyze the impact of missing nodes, edges, spurious nodes and edges, as well as other data artifacts, in networks with different properties, in a more theoretical and precise way.

3 Corporate Network Analysis & Data quality

In this section we present the corporate board networks obtained by interlocking directorates. An *interlocking directorate* occurs when a member of the board of an organization sits on the board of directors of another organization (Mizruchi, 1996). The original structure of the corporate system we are interested in, can be represented by a bipartite graph where the set of nodes is divided into two disjoint subsets: the *directors* subset and the *boards* subset (Caldarelli and Catanzaro, 2004). Every edge of the bipartite graph connects a director to a board, and vice-versa. If a director d_i sits on the board of a certain company c_j, an undirected edge from d_i to c_j will be present. Internal connections between nodes in the subsets are not allowed. This network can then easily be projected into two different one-mode networks: the *director network* where nodes are directors and weighted edges between them represent the companies they control. Second, the *corporate board interlock network*, which is the one we will focus on here. As previously mentioned, a corporate board interlock network can be defined as a network in which the nodes represent companies and the edges represent the board members these companies share.

We continue the rest of this section by discussing some of the most important work on corporate board network analysis and the study of centrality measures and community detection results. We then discuss data quality issues by referencing to some of the most important works regarding data quality and social network analysis.

3.1 Corporate Board Networks

Corporate board networks properties have been studied for several years.

Battiston and Catanzaro in 2004 show how the majority of the corporate networks are small world, have high average clustering coefficient, are assortative and they describe how the giant component covers most of the network.

In Windolf (2014) a study of seven German corporate networks from 1896 until 2010 was made. It focuses on the changes in density, position of the banks and intrasectoral networks, comparing the structure of the German corporate network with that of the United States. It finally shows how the density and the centrality of the banks tend to decrease over time, and for this, how an always stronger resemblance of the German network's structure

to that of the United States, can be observed.

Rinaldi and Vasta (2014) study the longitudinal behavior of the Italian corporate network. They take samples of the top 250 companies and their directors of the years 1913 to 2001, interpreting the evolution of the Italian corporate network of those years.

Croci and Grassi (2014) analyze the correlation between firm value and centrality measures, finding that degree and eigenvector centrality are negatively correlated with the revenue values. Grassi (2010) studies the topological structure of the Italian Stock Exchange corporate network and discusses the role of degree, betweenness and flow betweenness centrality in the network. It is suggested that usually hubs have a high degree, betweenness and flow-betweenness centrality. Companies having low degree and high betweenness usually aspire to be strategically connected in the network, while low degree and high flow-betweenness usually distinguishes banks.

A step further into understanding the role of the centrality measures in corporate networks has recently been done by Takes and Heemskerk (2016). In their work they provide a complete overview of the centrality measures investigating the global board interlock network consisting of circa $400,000$ companies connected by more than 1,500,000 shared board members. They then present *centrality persistence* and the *centrality ranking dominance*, where the first is able to quantify the persistence within the global network of the order of the most central firms of a single country, while the latter is able to compare rankings based on a partition and rankings based on the full global network.

In Piccardi et al. (2010) the analysis of the community structure of both the Italian board network and the Italian ownership network is presented. They show how an important division in communities on both networks is present. They then proceed by comparing the different community structures of the two networks by means of three set similarity measures: the rand index, the (normalized) van Dongen distance and the Variation of Information. Even though the two networks are technically distinct, all the set measures indicate a significant overlap between their partitions. Pyramidal groups (companies organized in layers) and strongly overlapping boards significantly contribute to the important (and similar) community structure of both networks.

Vitali and Battiston (2014) analyze the community structure of the global ownership network, built considering only transnational corporations (companies having the headquarter in one certain country and that operate in

at least one other foreign country). They start by unveiling the community structure of their global corporate network by means of the Louvain algorithm, to which they compare the community structure of rewired networks built from the same degree sequence and weights. The latter comparison highlights how the community structure of the empirical network is significantly different from the one of the rewired networks: the degree sequence and the weights are not enough to justify the community structure of their global ownership network. Vitali and Battiston then continue by investigating the existence of geographical and sectoral patterns in the largest eight communities of the giant component. They discover how geography plays an important role in the communities: the latter significantly reflect the geographical location of the companies. The sectors, instead, play only a secondary (and very marginal) role.

Finally, Heemskerk and Takes (2016) analyze the community structure of the global corporate interlock network built considering only "large and very large firms" which are "active" at this moment. Also, they merge all the firms of a country in one single node, focusing on the transnational interlocks. They show how running the Louvain community detection algorithm with resolution equal to 2 the Asian community immediately unveils. Asian counties are strongly connected within themselves and weakly connected with the rest of the world. Decreasing the resolution to 1.7 then (looking for subcommunities), they show how the Nordic and Baltic community appears. Lowering again the resolution to 1.5 and to 1.0, a Latin-American cluster and a western cluster (USA, Western Europe and the UK and Commonwealth) appear.

3.2 Data quality in networks

Costenbader and Valente (2003) analyze how eleven different centrality measures perform when random samples of directed networks (from 80% up to 10%) are taken. Networks of about 150 nodes are considered. Their results indicate that the *in-degree* centrality is, apart from the *eigenvector* centrality, the measure with higher correlation with the original network. They also show how the *out-degree* centrality decreases more rapidly, followed by the *closeness* and then *betweenness* centrality. Their results suggest that is possible to study networks generated from missing data.

One of the most well-known works on the topic is by Borgatti et al. (2006). Borgatti et al. present an extension of the work of Costenbader and Valente,

analyzing the robustness of degree, betweenness, eigenvector and closeness centrality under four types of error: node removal, edge removal, node addition and edge addition. In their work they show how all four centrality measures react surprisingly similar given a certain kind of error, and, for this reason, they suggest that an actual distinction between local and global centrality measures is not present — differently from what was previously thought. Unfortunately, despite the precious insights on the problem that the aforementioned works give, they both analyze very small sized networks. Networks of about 150 nodes and generated by an Erdős-Rényi random model (in Borgatti's case) do not resemble real-world networks features.

An important step to overcome these two problems has recently been done by Wang et al. In their paper Wang et al. (2012) analyze degree centrality, clustering coefficient, network constraint (see Burt (2009)) and degree centrality in six random error scenarios — the four presented by Borgatti et al. plus node splitting and merging — in two real-world networks with about $70,000$ nodes and $300,000$ edges. Their study suggests how in networks with positively-skewed degree distributions and high average clustering coefficients the four aforementioned network measures tend to be less resistant to errors. They also argue against the claim for which global measures are less resistant, but instead they suggest that the resistance of a certain measure to error scenarios can be associated with how it is actually computed.

4 Data

We extract samples of Italian, Danish, UK, Dutch, Spanish and Scandinavian companies from a 2013 snapshot of the Orbis — Bureau van Dijk dataset (Orbis — Bureau van Dijk, 2016). Each of the six samples (one for each country) will contain companies registered as "large" or "very large" and as "active". Also, only companies for which information about the senior directors was available, were selected. Then, interlocking the directorates in each of the six samples, we generate our corporate board interlock networks.

Even though we do not have a precise idea of how accurate and complete data in a certain country is, we believe that the selected countries, and in general most of the European ones, are more complete and accurate than the others.

The nature of the edges of corporate board interlock networks is typically weighted, where the weight represents the number of board members two companies share. For the sake of simplicity and following most of the literature in the field, here we will consider their unweighted (or binary) version. Eventually, we only analyze the giant components. The largest connected component, indeed, contains most of the information. Outside usually only lie a high number of small sized clusters of firms, of which the largest is much smaller than the giant component.

Finally, even though corporate datasets are particularly rich of information (such as the number of employees, revenue of companies, their geographical position and sector), here we will only take into consideration the revenue as a node attribute. In Figure 2, the Top 100 companies belonging to the Italian network's giant component are presented. Nodes are partitioned by their geographical location attribute. Eventually, from the two-mode network we remove all the edges (board members) whose positions are different from "board of directors", "executive board", "supervisory board" or "senior management".

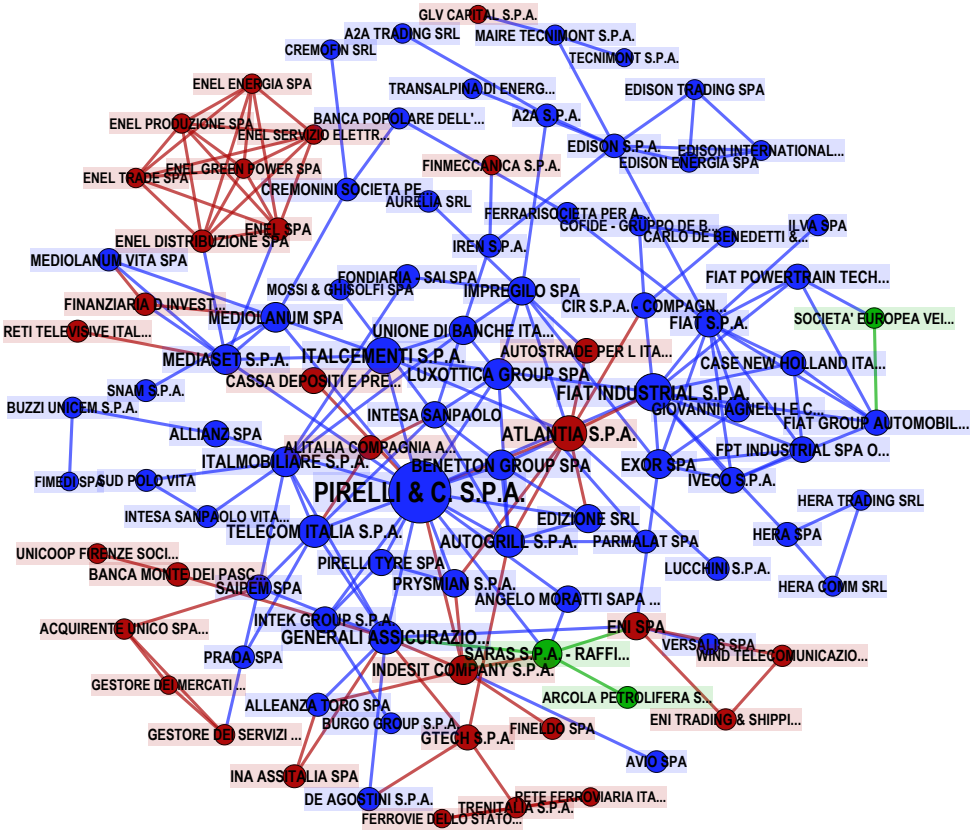


Figure 2: Top 100 nodes in the Italian network's giant component. Node size is proportional to betweenness centrality. In *blue* we represent companies in the north, in *red* central companies while in *green* we represent companies located in the south.

5 Methods

In the subsections that follow we present the fifteen data quality artifacts we use to stress our networks, in both a network science and a corporate network analytics context. We then proceed by discussing some details of the implementation and we finally conclude the section by presenting some of the measurements of the error we use to assess the effect of the artifacts. A general overview of the entire process in present in Figure 3.

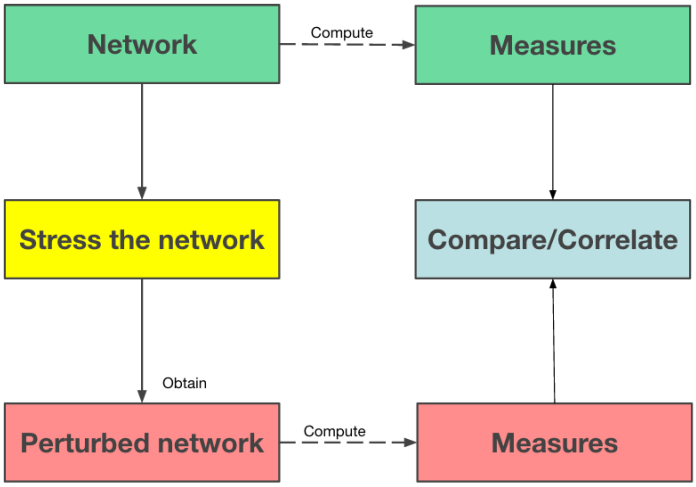


Figure 3: Overview of the process

5.1 Error scenarios

In order to reproduce some of the possible artifacts that one may encounter in a network, here we their test the resilience under fifteen different error scenarios. We start with the six scenarios defined in Wang et al. (2012), namely: node removal, edge removal, node addition, edge addition, node aggregation and eventually, node splitting. The procedures we use to replicate these errors, and their abbreviated names, are taken from the work by Wang et al. (2012). In order to follow their notation, from now on we will refer to these six error scenarios as follows:

- False negative nodes random,
- False negative edges random,

- False positive nodes random,
- False negative edges random,
- False aggregation random,
- False disaggregation random.

For the sake of clarity we decide to add the suffix "random" to the name of the scenarios. The reason behind this choice is the need for an easy and clear way to distinguish random and biased errors.

With the aim to let users and readers with a less technical background more easily understand our work, in Table 1 we present the artifacts we study in the simplest way possible. Throughout this work we will refer to the original network as G and to the "corrupted" (or "disrupted") one as H.

	Random	**Bias**
False Negative Nodes	We remove nodes. We simulate studying networks generated from datasets with missing companies.	Nodes with lower degree will be removed with higher probability. Usually, indeed, small companies are also the ones for which datasets have less information.
False Negative Edges	We remove board member's ties. We simulate a network built from a dataset in which board member's connections were missing.	Here we study what one sees when the missing ties connect mostly important (high degree) companies.
False Positive Nodes	We insert new nodes and we connect them with other ones already present. We assume to take a snapshot of the network in a point in time in which some percentages of companies, that should have been deleted, were not yet.	Here we simulate the situation in which (incorrect) shared board members ties are more likely to be present between spurious companies and other (already present) companies which have high degree.
False Positive Edges	We introduce new board member's ties to simulate studying a network in which shared board members that should have been removed, were not.	Here analyze the situation in which the network presents spurious shared board members ties, between at least one important (high degree) company over the two.
False Aggregation	Nodes merging. We simulate studying an old snapshot of a network in which companies that have split into two, are still present as a single company.	Here we assume that important companies are more likely to divide themselves into two companies. We reproduce this pattern aggregating with higher probabilities companies having high degrees.
False Disaggregation	Nodes splitting. We simulate the name of a company being spelled differently overall the dataset.	Companies for which the name is misspelled in the dataset are more likely to be having high degrees.

Table 1: The data quality artifacts.

In Chu and Davis (2011) and Heemskerk et al. (2016) examples of false negative nodes, false negative edges and false positive edges are presented. For an actual example of false disaggregation we instead refer to Chu and Davis (2015). More general considerations about quality of the data in the

Orbis dataset are presented in Heemskerk and Takes (2016).

Having to choose a measure on which we base our bias, we decide to use degree centrality. The reason behind this choice is that the degree is both an easily understandable measure in terms of corporate control (*power*) and it still resembles realistic biases that one may find in corporate datasets.

The formula we use to assign the probability $p \in [0, 1]$ of a node v to be selected from a vector $\vec{b}$ of length n, biased towards the degree, is presented in Equation 9:

$$b(\vec{v}) = (1 - \alpha)\frac{1}{n} + \alpha\frac{k_v^s}{\sum_{i=1}^{n} k_i^s} \tag{9}$$

where s is the slope of the degree distribution considered in logarithmic scale. In order to balance the bias more or less towards the degree, we introduce a constant parameter $\alpha \in [0, 1]$. With $\alpha = 0$ the bias will not be present — the nodes will be selected uniformly at random — with $\alpha = 1$, nodes with higher degree will have higher probabilities in the $\vec{b}$ vector. Eventually, given that in real-world networks the degree distribution follows a power law, and that in logarithmic scale this resembles a descending line with a certain slope coefficient s, here we decide to raise the degrees to the power of that coefficient. Doing this when building the $\vec{b}$ vector, we are taking into account the fact that nodes with low degree are much higher in number with respect to nodes with high degree and so their probability should also be much lower. Naturally, with $s = 1$, the assignment of the probabilities will be linear and plotting the degree distribution of the selected nodes we will have a power-law shape instead of the desired (almost) linear shape.

In order to better understand this, let us take the power-law degree distribution with slope $s = 2.5$ present in Figure 4 and let us plot the probability assigned to each degree in the case of $s = 1$ (see Figure 5a) and $s = 2.5$ (see Figure 5b).

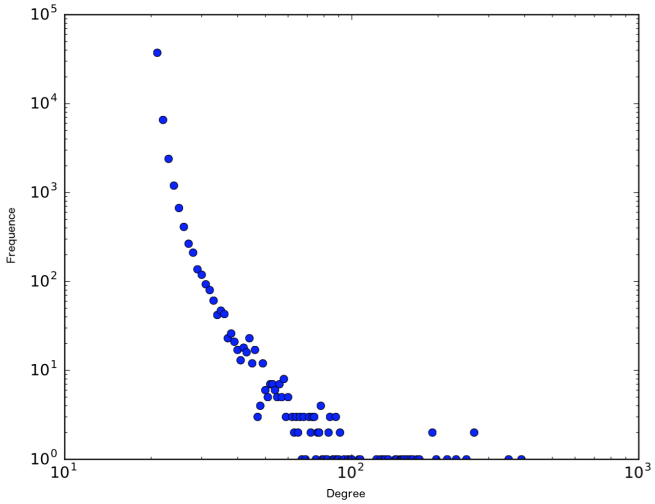
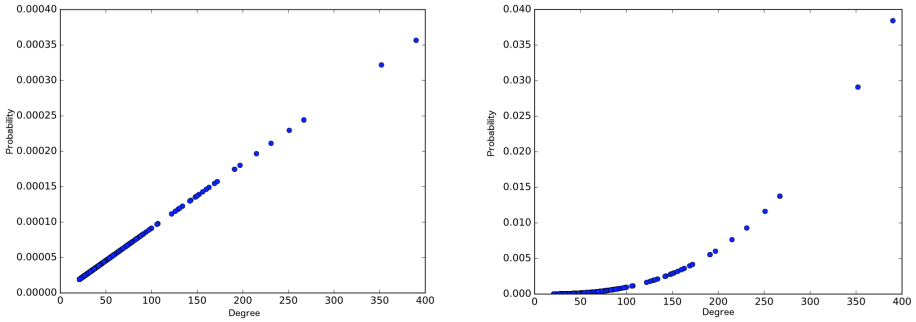


Figure 4: Power-law degree distribution in logarithmic scale

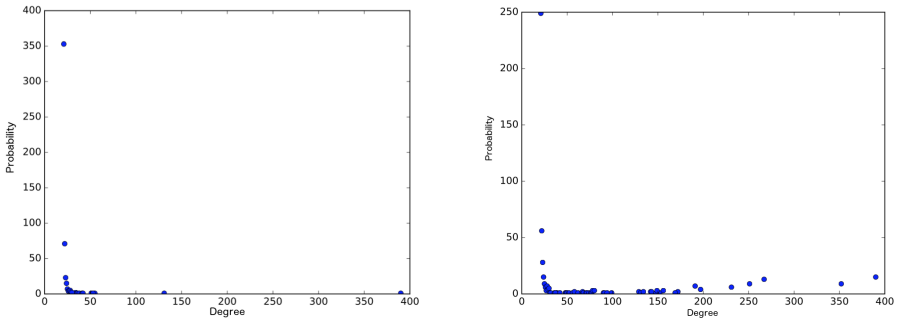


(a) (b)

Figure 5: Probabilities over degrees in case of $s = 1$ (a) and $s = 2.5$ (b)

As expected the probabilities in Figure 5a are distributed linearly, while the ones present in Figure 5b non-linearly, with an important increase in the right-most side of the plot. Selecting now 500 nodes with repetition following the two different probability laws, one can see how the degree distribution of the selected nodes changes. In particular, looking at Figure 6a we see how nodes with medium-high degree will never be selected, while if we also consider the slope of the curve in the computation of the probabilities, the

latter ones will be selected almost as often as medium-low degree nodes (see Figure 6b).



(a) (b)

Figure 6: Degree distributions of the selected 500 nodes at random in cases of $s = 1$ (a) and $s = 2.5$ (b)

In addition to the previous twelve artifacts we also want to study the consequences of artifacts biased towards one of the most typical firm properties: the revenue. We apply biased artifacts towards the revenue to all the error scenarios (see Table 2) apart from "false positive nodes", "false disaggregation" and "false aggregation". "False positive nodes" and "false disaggregation" expect the addition of spurious nodes for which the revenue would be impossible to estimate, while in false aggregation one should find a correct way to decide what revenue a new node, given the merging process of two other nodes, should have. Eventually, to asses the probability of each node to be selected with a bias towards the revenue, here we use the same formula in Equation 9, with $\alpha = 0.6$ and $s = 1$.

	Random	**Revenue**
False Negative Nodes	We remove nodes. We simulate studying networks generated from datasets with missing companies.	The probability that a node has to be selected is inversely proportional to its revenue. Negative revenues will be considered as zeros.
False Negative Edges	We remove board member's ties. We simulate a network built from a dataset in which board member's connections were missing.	Here we study what one sees if we remove with higher probability links between at least one wealthy company.
False Positive Edges	We introduce new board member's ties to simulate studying a network in which shared board members that should have been removed, were not.	Here we want to understand what do we see if we introduce with higher probability ties connecting at least one wealthy company (of the couple).

Table 2: The data quality artifacts biased towards the revenue.

5.2 Methodology

Here we explain the methodological details behind the artifacts we study.

For each of the fitness experiments we will have twelve iterations of experiments. At each iteration we start from the original network and we increase the error by 5%, eventually going from 5% in the first iteration to 60% in the last. The results of each iteration will then be averaged over ten runs, for a total of $15 \times 12 \times 10 = 1800$ runs. An important difference from the work of Wang et al. (2012) is that here we stop at an error rate of 60% — and not 95%. We believe, that an error of 95% percent is rather unrealistic, at least in modern corporate network datasets.

In the list that follows we present, for each artifact, the methodological procedure and its technicalities:

- **False Negative Nodes**
  - **Random**: We start each iteration by computing the number of nodes we need to remove. Let us call this number q. We then continue selecting q nodes at random, all together and without repetitions, and we remove them one at a time.
  - **Degree/Revenue**: We select the right number of nodes we need to keep, with a bias towards high degree (or wealthy) nodes. We then obtain the nodes to remove as the set difference between the entire list of the nodes and ones to keep.
- **False Negative Edges**
  - **Random**: We select a target node at random. If its degree is zero, we pick again. We then select at random one neighbor of the target node and we remove the edge that connects them. We continue until we have removed the desired number of edges.
  - **Degree/Revenue**: Here the target node and its neighbor are picked with a bias towards the degree (or the revenue).
- **False Positive Nodes**
  - **Random**: We start by adding the right percentage p of spurious disconnected nodes to the original graph. We then select at random p nodes, where the spurious ones are not included. We then start from the first random node, we compute its degree d and we

eventually connect one edge from the first spurious node to other d random ones. We continue until each disconnected spurious node has been connected.

- **Degree**: We add p spurious disconnected nodes. We then select at random p nodes. The spurious ones are not included. We then start from the first random node, we compute its degree d and we eventually connect one edge from the first spurious node to other d selected with a bias towards high degree ones.

- **False Positive Edges**

  - **Random**: We start by selecting two random nodes. If they are already connected, we pick another couple, otherwise we connect them with an edge.
  - **Degree/Revenue**: We start by selecting a node v_1 with a bias towards the degree (or the revenue). We then select another node v_2, this time uniformly at random, with v_1 and v_2 not connected and we proceed by connecting them.

- **False Aggregation**

  - **Random**: We start by selecting two different random nodes. The first node v_1 will be the node to maintain while the second node v_2 will be the one that will be merged into the first. We then proceed by attaching the neighbors of v_2 to v_1 and by deleting v_2. We continue until the right number of nodes have been merged.
  - **Biased**: We select the right percentage p of nodes we need to keep, with bias towards the degree. We continue by selecting other p nodes with bias, which will be the ones to remove. We then merge the nodes as before, but this time, considering one node couple of nodes at time: one node to keep and one to remove.

- **False Disaggregation**

  - **Random**: We start by randomly selecting one node to split at a time. After having selected the node, we list its neighbors and we create a new spurious disconnected node. We then proceed by attaching, at random, 50% of the edges of the node to split to the new spurious node, as presented in Wang et al. (2012). We

continue this process until the right number of nodes have been split.

- **Biased**: We select one node to split at time, with biases towards the degree. After having selected the node, we list its neighbors and we create a new spurious disconnected node. We then proceed by attaching at random 50% of the edges of the first node (the one to split) to the new spurious node.

5.3 Implementation

The code is run on a server with 16 Intel Xeon E5-2630v3 CPUs @ 2.40GHz (32 threads) and 1.5TB of RAM. It has been entirely written in Python using the graph-tool library (Peixoto, 2014), whose core data structures and algorithms are implemented in C++. The reason behind this choice are both the higher readability Python has with respect to Java and C++ and the extremely optimized and parallelized functions that the graph-tool library guarantees.

To have better performance, in the implementation of each error scenario we decide to store the twelve graphs in a compressed format and to then execute the Louvain algorithm and the variation of information in parallel using one graph for each core.

5.4 Measurements of the error

In order to understand how H differs from G, we study the changes in some of the most important network metrics and distributions. The metrics we study are presented in the list that follows:

- Degree distribution
- Distance distribution
- Average neighbor correlation
- Percentage of nodes and edges in the giant component
- Density
- Global clustering coefficient

- Average distance (on the network's giant component)
- Scalar assortativity coefficient

Apart from the aforementioned metrics and distributions, we are also interested in studying how degree centrality, betweenness centrality and closeness (harmonic) centrality results change when increasing the error rate. To do so we follow the work of Wang et al. (2012): we start from our original graph G and we introduce the desired error percentage, obtaining the "corrupted" graph H. We then take the set of companies that G and H have in common and we call this set of nodes $C = V^G \cap V^H$. Eventually, for each node $v \in C$ we compute its degree centrality, betweenness centrality and harmonic centrality for both graphs G and H. We save the result of company v for graph G in $\vec{z}$ and the result of company v for H in $\vec{z'}$. Finally, once all the results for each company in C have been stored, we compute the Spearman's ρ ranking correlation coefficient between $\vec{z}$ and $\vec{z'}$.

5.4.1 Spearman's rho vs Kendall's tau

When measuring the relation between different rankings of the same (or different) variable, the following two ranking coefficients are most frequently used: Spearman's rank correlation (ρ) and Kendall's rank correlation (τ). The general idea is simple: in both measures, if the two rankings are equal, the value of both correlation coefficients will be 1. In case the rankings are completely different the correlation coefficient will be 0. In case one ranking is the exact reverse of the other, both correlation coefficients will be -1. For more information we refer the reader to Langville and Meyer (2012).

The substantial difference between these two coefficients is in the different importance they give to sequential swaps in the rankings. More specifically, Kendall's τ gives the same importance (penalty) to both small and big drops (or gain) in the ranking while Spearman's ρ gives higher penalties to the latter ones. In other words, when comparing rankings one should use Spearman's ρ if the number of positions a certain value has lost (or gained) in the ranking is of importance. If only the number of concordant and non-concordant pairs matters than the Kendall's τ is the one to use.

Given that we are particularly interested in penalizing the leaps (number of positions lost or gained) that companies make in the "importance" ranking when artifacts in the network are applied, the Spearman's ρ correlation coefficient is the most appropriate measure for us to use. Indeed, we are not only

interested in understanding if a company in G has, for instance, different degree centrality than in graph H, but we are also interested in understanding (and penalizing) the number of positions it has lost or gained.

Rank	Ranking in G	Ranking in H
1st	A	H
2nd	B	B
3rd	C	C
4th	D	D
5th	E	E
6th	F	F
7th	G	G
8th	H	A

Table 3: Ranking in the original graph G versus the ranking in the "corrupted" graph H

Let us take for instance the two rankings in Table 3. One can see how the only difference in the two rankings is the inverted position of the first and last company. In this case Kendall's τ gives a 0.071 correlation value (meaning that the two rankings are different), Spearman's ρ gives a correlation value of -0.166. The latter value, even in this very simple example, has greater tendency to indicate the two rankings as inverted. The number of positions company A has lost, and the number of positions company H has gained, has been penalized more in the latter correlation coefficient. To understand how the structure of H has changed with respect to G, penalizing big leaps is of extreme importance.

5.4.2 Variation of information

To understand how the community structure changes, we compute the variation of information (Meilă, 2007) on Louvain's partitions. In particular, for each node $v \in G \cap H$, we take its community number and we store it in vectors $\vec{V}$ and $\vec{V'}$, respectively. Finally, we compute the variation of information between vectors $\vec{V}$ and $\vec{V'}$.

The variation of information is a metric in the space of partitions, elaborated by Meilă (2007). It lies in the range $[0, \log n]$, thus we can simply normalize it in the range $[0, 1]$ dividing it by $\log n$.

The variation of information will be maximal when the two partitions are completely different, namely: X tells nothing about Y and vice-versa, and so the mutual information between X and Y is zero. In the same way, the variation of information will be zero when the two partitions are exactly equal. Also, it is a true metric on the space of clusterings (see Meilă (2007)), meaning that is non-negative, symmetric and it satisfies the triangle inequality. For this reason, and following other related work, such as Piccardi et al. (2010) and Vitali and Battiston (2014), we decide to adopt this measure over other indexes of comparison.

5.4.3 Kolmogorov-Smirnoff Two Sample Test

To measure the differences between the degree and distance distributions of the graph G with those of the perturbed graphs H, here we use the Kolmogorov-Smirnoff (KS) two sample test. This is a two-sided test meant to assesses whether two independent samples are drawn from the same distribution. Hence, given two distributions one can test for a null hypothesis, i.e., the two samples are drawn from the same distribution, inferring on the so called *D-statistic* and *p-value*. The general idea is the following: if the *p-value* is small and the *D-Statistic* relatively high one can state that the two populations were sampled from different distributions.

Despite its main goal, here we decide to do not use the KS test to assess whether two independent samples are drawn from the same distribution, but only as a measure of difference: the maximal difference between the two cumulative probability distributions of the two populations. The reason behind this choice is the nature of our degree and distance distributions, which are *discrete*. The assumption behind the KS test, indeed, is that the two distributions compared are *continuous* and so, without ties. If used on discrete distributions the result might be misleading.

Another usable measure would have been the so-called Chi-Square Goodness of Fit Test. Despite being able to work with discrete distributions, the Chi-Square Goodness of Fit Test it is still not suitable in our case: the expected number of observations in each level of the variable must be at least five and the number of bins must be equal. Given the power-law degree distributions we encounter, both prerequisites are at least difficult to satisfy.

Even though some non-parametric tests meant for continuous distributions have also been adapted for discrete distributions, as mentioned in Arnold and Emerson (2011), here we decide to avoid this path, but instead

we simply prefer to use the *D-statistic* as a metric. Finally, even though the *Chi-Square Goodness of Fit Test*, with some adjustments, might have been (theoretically) usable for the distance distributions, given the number of samples in each bin and the simplicity of the *KS* test, we decide to adopt the latter test for both distributions.

6 Experiments

In this section we present the impact of imperfect data in real-world corporate networks and we propose a simple and effective way to visualize and understand how resilient a network is.

In Section 6.1 we present detailed results for the Italian corporate network's giant component while in Section 6.2 we discuss the results of the Italian, Danish, Great Britain, Dutch, Spanish and Scandinavian network's giant components by means of what we called a *resilience matrix*.

For reasons of space and comprehensibility in all the subsections that follow we only present the most significant figures.

6.1 The Italian corporate network's giant component

Here we study the resilience of the Italian corporate network's giant component, stressing the network with the error scenarios presented in Table 1 and Table 2. In Section 6.1.1 we first present the topological properties of the network under study.

6.1.1 Network properties

The visualization of the Italian corporate board interlock network's giant component is presented in Figure 7. The number of cliques and high density clusters is relatively low with respect to the overall size of the network, which has $4,483$ nodes and $12,517$ edges, letting us believe that the quality of the data is at least sufficient for the study.

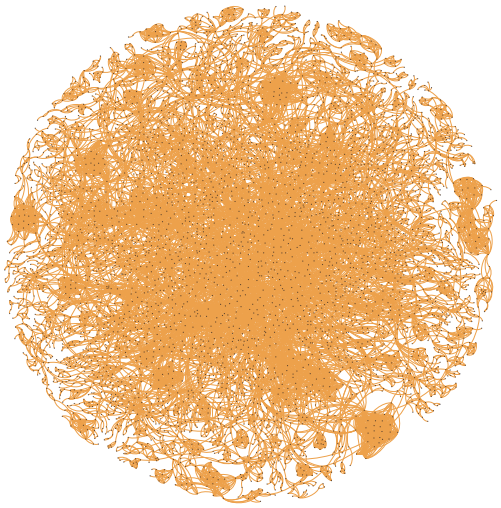
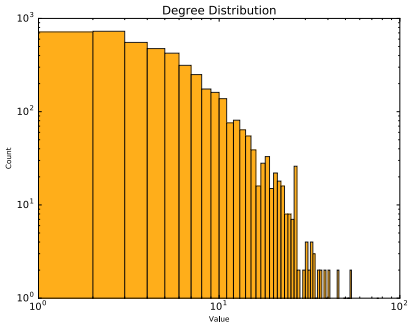
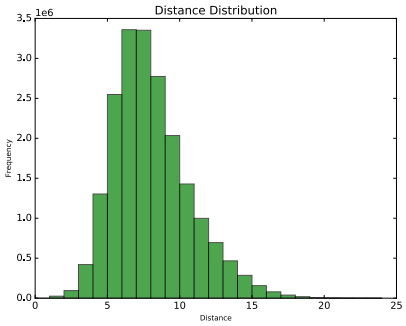


Figure 7: Visualization of the Italian corporate board interlock network's giant component. Visualized using Gephi (Bastian et al., 2009) and ForceAltas 2 layout with "Stronger Gravity" enabled and "Scaling" coefficient set to 30.

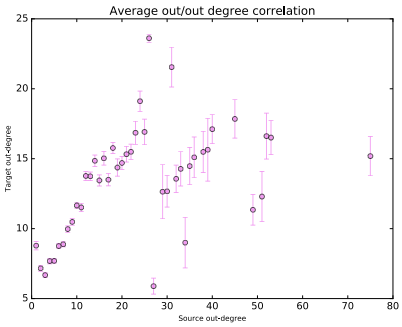
We proceed by looking at its degree distribution, distance distribution and its average neighbor correlation, presented in Figures 8a, 8b and 8c, respectively.



(a) Degree distribution



(b) Distance distribution



(c) Average neighbor correlation

Figure 8: Degree distribution, distance distribution and average neighbor correlation of the Italian corporate network's giant component.

The density, global clustering coefficient, average distance and assortativity coefficient are finally presented in Table 4.

Measure	Value
Density	0.0012
Global clustering coeff.	0.524
Avg distance	7.569
Assortativity coeff.	0.296

Table 4: Topological properties of the network. The colors on the left of each measure represent the relative colors we will use in the plots that will follow.

Looking at the measures and distributions present, we see how the Italian

network's giant component characteristics reflects the ones found in most real world ones: the degree distribution follows a power-law and it is a small-world network. In the subsections below we now present the detailed changes of all the measures under the 15 error scenarios.

6.1.2 False negative nodes - Random

The global measures never change significantly. In particular, the density remains stable with only negligible fluctuations. The same happens to the global clustering coefficient and to the average distance. The latter can also be observed looking at the distance distribution changes in Figure 9b, where the vertical axis (frequency) decrease homogeneously. Eventually, the assortativity coefficient is the only one that stays pretty stable up to the 45% of error, point where we register a single fluctuation of about 10%. From then on it restores and keeps is initial value.

The situation seems to be clear: removing nodes at random, the probability that the selected nodes have low degrees is much higher than having selected nodes with high degrees. To prove it we do the following: knowing the number of nodes and edges of graph G (to which we will refer as n and m) and the number of nodes and edges of the perturbed graphs H (n' and m', respectively), we can easily compute the average degree of the removed nodes (average number of connections), doing

$$\frac{m - m'}{n - n'}$$

which is around 4 at each error rate. This clearly happens because of the power-law degree distribution: the number of nodes with low degrees is much higher than the number of nodes with high degrees. Now, having in mind that removing nodes also implies removing edges and so reducing the degrees of the neighbors of the removed nodes, here we see a "scale reduction" effect.

The density considered on the giant component of the perturbed networks increases. This is due to the continuous reduction of the giant component size in which mostly medium-low degree nodes have been removed. This entails that the remaining nodes in the giant component will likely have relatively high degree, although they will only be a few in number.

The global clustering coefficient stays almost untouched, likely due to the fact that low degree nodes do not participate in any triangle, or at least in relatively few over the total number. The higher the degree, the higher is

the probability that you are part of one or more triangles.

These ideas are also confirmed by looking at how the degree distribution changes over time (Figure 9a). We see a natural decrease in the vertical axis (frequency) but just a slight difference in the horizontal axis (degree), indicating that the likelihood that, even after removing 60% of the nodes at random, the high degree nodes in G will still have relatively high degree in H is significant.

Eventually, the Spearman correlation on the three centrality measures (see Figure 10), we notice a good robustness.

Of the three measures considered, degree centrality seems to be most stable, followed by betweenness and then by harmonic centrality. We attribute this behavior to the nature of the three different measures. Each time a random node is removed from the network, the degree of all its neighbors decreases just by one. Moreover, since nodes with lower degrees have higher chances to be removed at random, this only influences a relatively small number of other nodes (neighbors of the selected one). This means that only relatively few nodes, each time one of their neighbors is removed, will see their degree centrality lowered. Also, each time a node is removed, the degree of its neighbor can be lowered at most by one. These considerations let us think that degree centrality it is actually a pretty robust measure in itself under this data quality artifact.

When removing random nodes the shortest paths can only become longer or untouched, but not shorter. Imagine having two paths of different length from one node v_1 to another node v_2. Now, if we select a node in the shortest path between v_1 and v_2, the length of the path will be enlarged (what before was the longest path is now the shortest) while if we remove a node in the longest path we are actually leaving the shortest one of the same size. In the worst case the longest path will be infinite, meaning that there will no longer be edges from v_1 to v_2.

In general, nodes with degree one will have a betweenness centrality value of zero, while the same is not true for harmonic centrality. Unconnected nodes will have zero betweenness centrality as well as zero harmonic centrality. One of the reasons behind the results presented in Figure 10 might be that removing nodes at random carries a heavy disconnection of the graph. This will entail an always higher number of nodes with betweenness centrality zero and harmonic centrality zero, which will negatively influence the results of the Spearman correlation. The main reason for which betweenness centrality (for this network, error type and apart from the aforementioned problem)

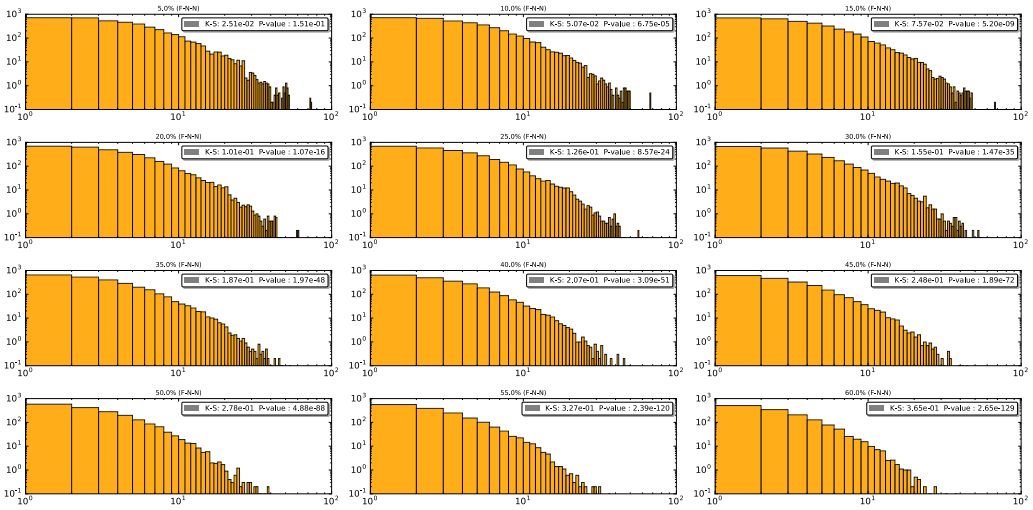
maintains a better ranking correlation than harmonic centrality will be explained in Section 6.1.3, after the results for the false negative nodes biased have been presented.

6.1.3 False negative nodes - Degree bias

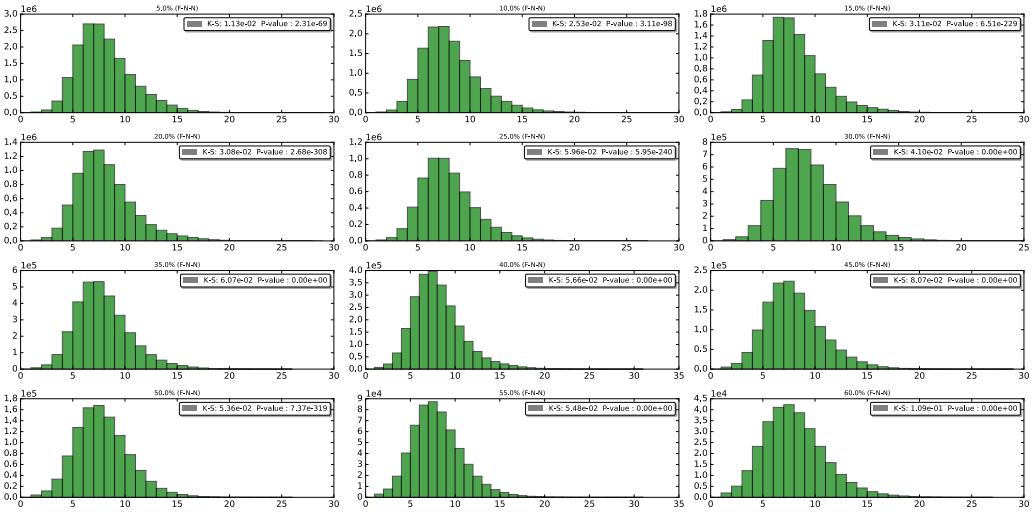
Until now we have remarked how "false negative nodes" is in some sense biased towards low degree nodes, due to the power-law degree distribution. But how biased is it? What if we stress the bias? And how the network react to a stronger bias? In order to answer these questions we introduce "false negative nodes biased". The results suggest that the "scale reduction" effect we obtain in the previous error scenario here disappears. Setting an $\alpha = 0.7$ in Equation 9 we see how this biased version behaves more aggressively on the graph. In particular, looking at Figure 11 we notice a significant increase of the density up to 140% (from 0.012 to 0.030), an increase of the global clustering coefficient and the assortativity coefficient up to almost 18% and 50% respectively. We clearly see how removing high numbers of low degree nodes, the average degree increases. The number of links does not decrease as fast as the number of nodes (increasing the density). The number of triangles decreases less rapidly than the number of connected triplets of nodes (increasing the global clustering coefficient). The number of low degree nodes dramatically decreases but only partially diminishing the degree of the rest of the nodes (increasing the assortativity coefficient) and the perturbed graphs are much less disconnected than in the uniformly random case (see Table 7). Given all of this, we finally notice a decrease in the average distance.

Another interesting result is the different behavior of the giant component with respect to "false negative nodes at random". Looking at Table 6 we see how the density of the giant component in the aforementioned error scenario increases more than the density in the case of biased artifact. If in the latter error scenario the giant component at 60% error has about $1,200$ nodes with more than $4,000$ edges, in the random case the giant component has a comprehensively smaller size (about 500 node and 1000 edges). In general the more a (real-world) network increases in number of nodes the more difficult it will be for the density to increase (see the Facebook example in Section 2.1).

For what concerns the three centrality measures we also obtain interesting results: the degree and harmonic centrality behave practically the same, reaching a correlation of almost 0.9, while betweenness centrality decreases



(a) Degree distribution



(b) Distance distribution

Figure 9: Degree distribution, distance distribution and average neighbor correlation of the Italian corporate network's giant component in false negative nodes for each error percentage.

up to 0.78. The different behavior of harmonic centrality, with respect to the random error may be explained by the lower number of disconnected com-

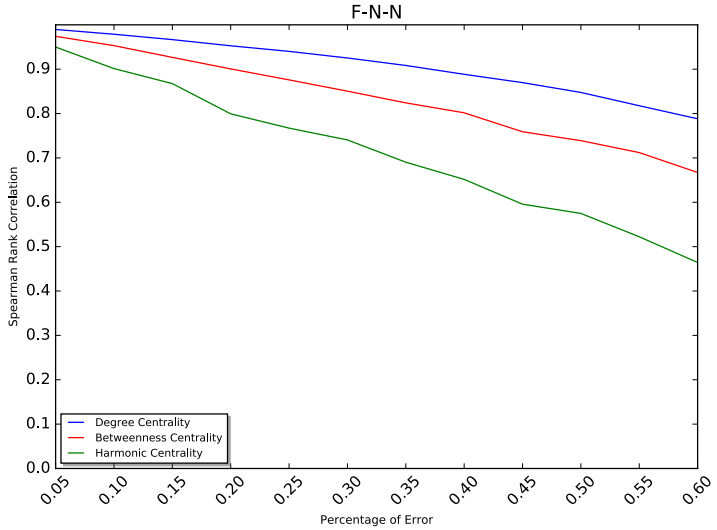


Figure 10: Spearman correlation of degree centrality, betweenness centrality and harmonic centrality for false negative nodes at random.

ponents. In particular, for 60% error here we obtain 26 different components with respect to the 60 we obtained in the random case. One may try to attribute the different behavior of the harmonic and betweenness centrality to the different topology of the perturbed graphs under the two different types of errors. In particular, looking at Table 5 one can see how the random error disconnects the graph in more homogeneous (in size) connected components, while the biased error maintains a large giant component surrounded by many components of smaller sizes. This result lets us imagine why harmonic centrality in one case (random, see Figure 10) descends rapidly and more than betweenness centrality, while in the second case it remains almost always stable up to 0.9. More concretely: computing harmonic centrality in the biased case does not alter the ranking much, since the majority of the nodes will see their rankings lowered by the same number of zeros (given by the disconnected nodes). The same does not happen in the random error scenario where, given the more homogeneous disruption of the network, many of the nodes in all of the components will see their ranking change.

	Nodes	Edges	Nodes Gc	Edges Gc	# of components
FNN	1794	1986	575 (32%)	1059 (53%)	26
FNNB	1794	4836	1254 (69%)	4278 (88%)	60

Table 5: Size of the perturbed networks under false negative nodes random and biased at 60% of error. The percentages reported are relative to the first and second column of the table.

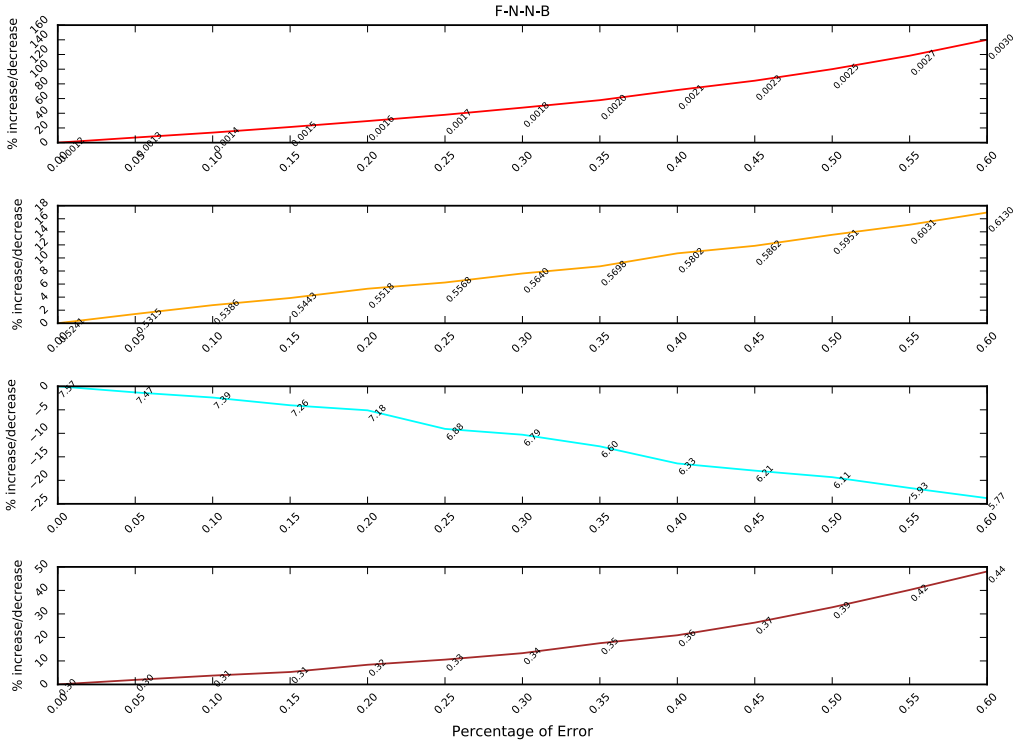


Figure 11: Measure changes under false negative nodes biased for each error rate.

6.1.4 False negative nodes - Revenue bias

Here nodes are removed inversely proportional to their revenue, with the parameter $\alpha = 0.6$. Knowing that there is usually a weak correlation from revenue to degree in corporate networks, we expect a "trade-off" error between "false negative nodes at random" and "false negative nodes biased towards the degree". Looking at the results of the global measures we see how the

assortativity coefficient, apart from some initial swings, tends to remain the same. The density raises up to +35%, the clustering coefficient decreases up to −25% with a 60% error. The same happens to the average distance. With respect to the previous two errors, here the network reacts more or less as expected, and no particular peaks or changes are observed.

The same happens to the Spearman correlation: the values and the tendencies of the curves are in the middle of the ones seen in "false negative bias" and in its "random case". The degree centrality decreases up to about 0.83, while betweenness centrality and harmonic centrality decrease up to 0.75. Eventually, also the number of components and the density of the giant component values (see Table 7 and Table 6) are typically in the middle between the random and biased cases.

	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**
FNNR	0.0013	0.0014	0.0016	0.0017	0.0019	0.0021	0.0023	0.0026	0.0030	0.0034	0.0040	0.0046
FPER	0.0013	0.0014	0.0014	0.0015	0.0016	0.0016	0.0017	0.0017	0.0018	0.0019	0.0019	0.0020
FNER	0.0013	0.0013	0.0013	0.0014	0.0015	0.0015	0.0017	0.0018	0.0019	0.0021	0.0023	0.0026
FNN	0.0014	0.0015	0.0016	0.0018	0.0020	0.0023	0.0026	0.0029	0.0036	0.0040	0.0051	0.0068
FPN	0.0012	0.0012	0.0012	0.0013	0.0012	0.0012	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013
FNE	0.0013	0.0013	0.0014	0.0015	0.0016	0.0017	0.0018	0.0021	0.0022	0.0024	0.0027	0.0031
FPE	0.0013	0.0014	0.0014	0.0015	0.0016	0.0016	0.0017	0.0017	0.0018	0.0019	0.0019	0.0020
FA	0.0014	0.0015	0.0017	0.0019	0.0022	0.0025	0.0029	0.0034	0.0041	0.0049	0.0061	0.0077
FD	0.0012	0.0012	0.0011	0.0011	0.0011	0.0011	0.0011	0.0010	0.0010	0.0010	0.0010	0.0010
FNNB	0.0014	0.0015	0.0017	0.0018	0.0020	0.0023	0.0026	0.0030	0.0034	0.0039	0.0046	0.0054
FPNB	0.0012	0.0012	0.0012	0.0012	0.0012	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013
FNEB	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0013	0.0013	0.0014
FPEB	0.0013	0.0014	0.0014	0.0015	0.0016	0.0016	0.0017	0.0017	0.0018	0.0019	0.0019	0.0020
FAB	0.0014	0.0015	0.0017	0.0019	0.0022	0.0025	0.0029	0.0034	0.0041	0.0050	0.0061	0.0077
FDB	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009	0.0009	0.0009	0.0008

Table 6: Density values in the giant component of the network for every artifact and error rate.

	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**
FNN	7	13	22	26	33	36	46	43	57	54	58	61
FNNB	3	5	8	11	13	17	19	20	23	25	22	27
FNNR	5	12	12	18	22	31	32	35	39	40	41	42
FNE	13	33	47	62	84	105	127	145	169	192	210	236
FNEB	5	14	18	26	37	46	52	63	77	85	99	120
FNER	9	21	36	49	66	83	103	123	144	163	182	213
FD	14	30	45	64	77	96	113	129	145	169	179	199
FDB	12	24	38	48	64	78	92	106	119	130	141	152

Table 7: Number of connected components in the perturbed networks at each error rate.

6.1.5 False negative edges - Random

Let us start by pointing out that where removing nodes uniformly at random means, in some senses, that we have a bias towards low degree nodes, removing edges at random we have a different bias. In particular, when we remove nodes at random we have higher probability to have selected a low degree node, while if we select an edge at random we have a high probability to have selected an edge belonging to at least a node (of the pair) with high degree.

Looking at Figure 12 we notice how the properties of the graph change with respect to "false negative nodes". If the number of nodes remains constant, the number of edges (and so the density in the perturbed graphs H) drops proportionally to the error rate. The clustering coefficient decreases as expected, while the average distance very slowly descends from 7.57 at 0% to 6.64 at 60% of error. If one would expect the average distance (of the giant component) to increase given the removal of edges from high degree nodes, here we witness the opposite behavior. Eventually the assortativity coefficient increases by a negligible 14%, going from 0.30 up to 0.34 at 60% of error.

Apart from degree centrality that decreases up to 0.8, the harmonic and betweenness centrality are significantly less stable than before, reaching correlations of 0.55 and circa 0.43, respectively, at 60% of error (see Figure 13a). The surprising result here is that, even though the perturbed networks H count a much higher number of disconnected components with respect to "false negative nodes at random" (see Table 7), harmonic centrality seems to be more stable than betweenness centrality (the opposite happened removing nodes at random).

The degree and distance distributions, as well as the average neighbor correlation, do not undergo surprising changes, or at least noteworthy, changes. In general though, "false negative nodes at random" seems to be much less stressful for the network compared to "false negative edges at random" (see Table 8).

6.1.6 False negative edges - Degree bias

Reminding that "false negative edges at random" is biased towards high degree nodes, here we want two answer the same two questions we asked before, namely: "how much biased is it?" And "what if we stress the bias?" In order

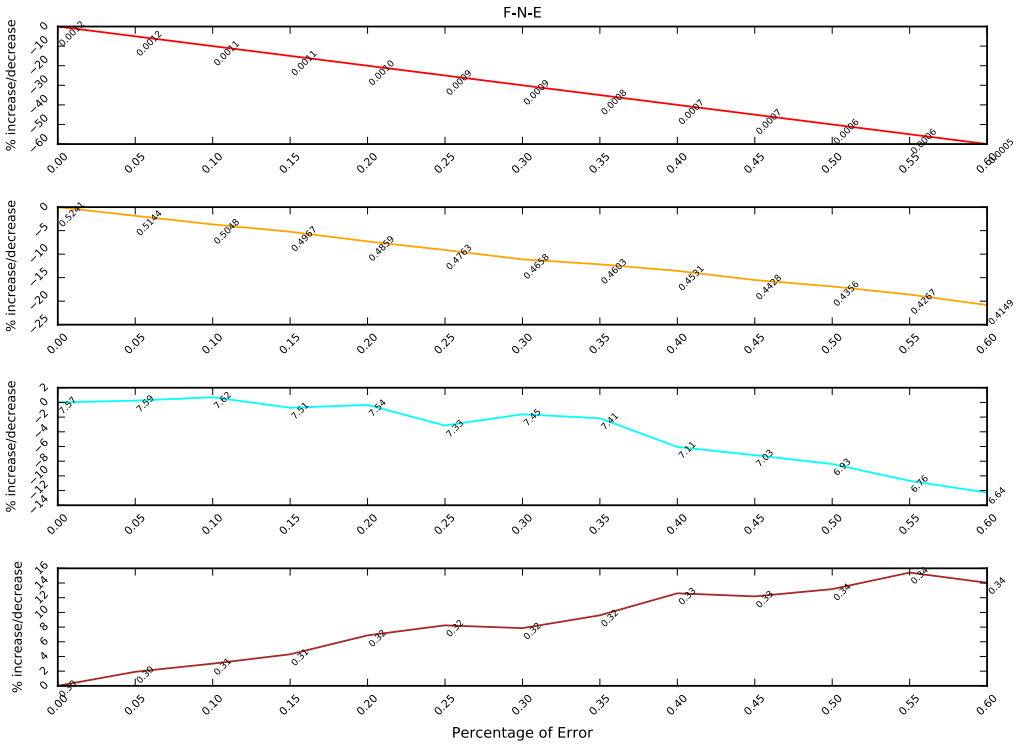
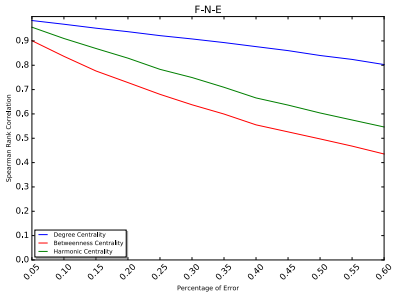


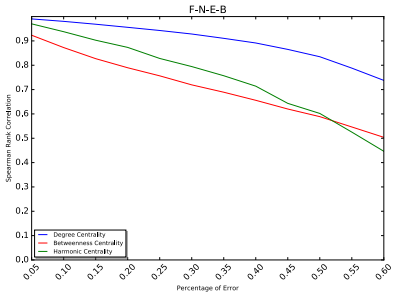
Figure 12: Measure changes under false negative edges at random for each error rate.

to answer these questions here we analyze the results of "false negative nodes biased towards high degree nodes".

Starting from Figure 13b we see how the betweenness and closeness ranking correlations are close to each other. Despite this, the situations for the global measures happens to be a bit different. In particular, looking at Figure 15, if the density and the global clustering coefficient descend almost as much as in the random case, the average distance and the assortativity coefficient completely change trend. The average distance (in the giant component) increases up to a remarkable 70%, going from 7.57 to 12.59 at 60% of error. With respect to Figure 12 where a descend in the average distance was present, here the ascent naturally makes more sense: removing edges mostly from very high degree nodes, a high number of shortest paths will likely be cut from the network, and reaching two non-adjacent vertices (following shortest paths) will naturally take more steps. Now, the difference of



(a) Spearman correlations for false negative edges at random.



(b) Spearman correlations for false negative edges biased.

Figure 13: Spearman correlation of degree centrality, betweenness centrality and harmonic centrality for false negative edges at random and false negative edges biased.

the average distance in the two error scenarios might be understood by looking at the densities and the sizes of the giant components (for convenience at 60% of error), presented in Table 9. If for the random case the perturbed network has almost 40% of the nodes and almost all the edges left in the graph (with a density of 0.0031), in the biased error we have almost 50% of edges with 70% of the nodes left. The density is equal to 0.0014, which is almost the half. In other words, the number of possible shortest paths in the perturbed network under "false negative edges at random" is much higher than the number of possible shortest paths of the perturbed network under the "false negative edges biased" error scenario. A better view of this phenomenon can be observed by looking at Figure 14a and Figure 14b where one can see how in the random scenario the horizontal axis (distance) decreases when increasing the error rate, while in the biased scenario the distances increases.

Another significant change is the decrease of the assortativity coefficient from an error rate of circa $35 - 40\%$. When the error rate reaches 60% the assortativity coefficient decreases by almost the 35%, going from 0.30 to 0.19. The network is becoming increasingly more non-assortative. A better view of this phenomenon can be seen by looking at Figure 16 where the line is becoming more horizontal. There is less correlation between the degree of a node and the degree of its neighbors.

	Deg. Centr.	Betw. Centr.	Harm. Centr.	Density (%)	Global c.c. (%)	Avg. Dist. (%)	Assort. coeff. (%)	Var. Info.	KS Deg	KS Dist	Nodes GC	Edges GC
FNNR	0.83	0.74	0.77	0.0017 (37.67%)	0.4008 (-23.52%)	5.89 (-22.24%)	0.28 (-4.88%)	0.30	0.27	0.28	970 (21.64%)	2174 (17.37%)
FPER	0.89	0.48	0.31	0.0020 (60.00%)	0.1110 (-78.82%)	3.56 (-53.01%)	-0.04 (-112.00%)	0.35	0.23	0.82	4483 (100.00%)	20027 (160.00%)
FNER	0.72	0.38	0.39	0.0005 (-60.00%)	0.5246 (0.09%)	9.09 (20.12%)	0.57 (91.98%)	0.38	0.44	0.23	1708 (38.10%)	3834 (30.63%)
FNN	0.79	0.67	0.46	0.0012 (-1.18%)	0.5264 (0.43%)	7.44 (-1.68%)	0.33 (10.19%)	0.32	0.36	0.11	542 (12.09%)	1002 (8.01%)
FPN	0.83	0.46	0.51	0.0013 (3.31%)	0.1648 (-68.55%)	4.29 (-43.31%)	0.18 (-38.92%)	0.47	0.39	0.75	7172 (159.98%)	33098 (264.42%)
FNE	0.80	0.44	0.55	0.0005 (-60.00%)	0.4149 (-20.83%)	6.64 (-12.28%)	0.34 (14.03%)	0.39	0.50	0.14	1675 (37.36%)	4320 (34.51%)
FPE	0.86	0.52	0.59	0.0020 (60.00%)	0.2815 (-46.29%)	4.34 (-42.71%)	0.28 (-5.01%)	0.42	0.38	0.74	4483 (100.00%)	20027 (160.00%)
FA	0.46	0.34	0.27	0.0077 (516.60%)	0.2164 (-58.71%)	3.26 (-56.99%)	0.05 (-83.01%)	0.69	0.35	0.87	1794 (40.02%)	12355 (98.71%)
FD	0.95	0.83	0.89	0.0005 (-60.93%)	0.4608 (-12.08%)	7.94 (4.96%)	0.28 (-6.48%)	0.13	0.29	0.05	5009 (111.73%)	12294 (98.22%)
FNNB	0.92	0.77	0.90	0.0030 (140.03%)	0.6130 (16.96%)	5.77 (-23.79%)	0.44 (48.05%)	0.23	0.12	0.31	1258 (28.06%)	4257 (34.01%)
FPNB	0.94	0.60	0.54	0.0013 (3.45%)	0.0935 (-82.15%)	3.95 (-47.87%)	0.03 (-90.22%)	0.40	0.18	0.76	7172 (159.98%)	33144 (264.79%)
FNEB	0.74	0.50	0.45	0.0005 (-60.00%)	0.2982 (-43.10%)	12.59 (66.34%)	0.19 (-34.40%)	0.33	0.32	0.56	2226 (49.65%)	3502 (27.98%)
FPEB	0.95	0.62	0.53	0.0020 (60.00%)	0.1678 (-67.98%)	3.97 (-47.57%)	0.10 (-66.38%)	0.32	0.17	0.75	4483 (100.00%)	20027 (160.00%)
FAB	0.14	0.10	0.07	0.0077 (519.04%)	0.2396 (-54.27%)	3.30 (-56.36%)	0.12 (-59.06%)	0.78	0.41	0.86	1794 (40.02%)	12404 (99.10%)
FDB	0.95	0.81	0.88	0.0005 (-60.93%)	0.3914 (-25.32%)	8.50 (12.34%)	0.23 (-23.71%)	0.16	0.25	0.14	5403 (120.52%)	12156 (97.12%)

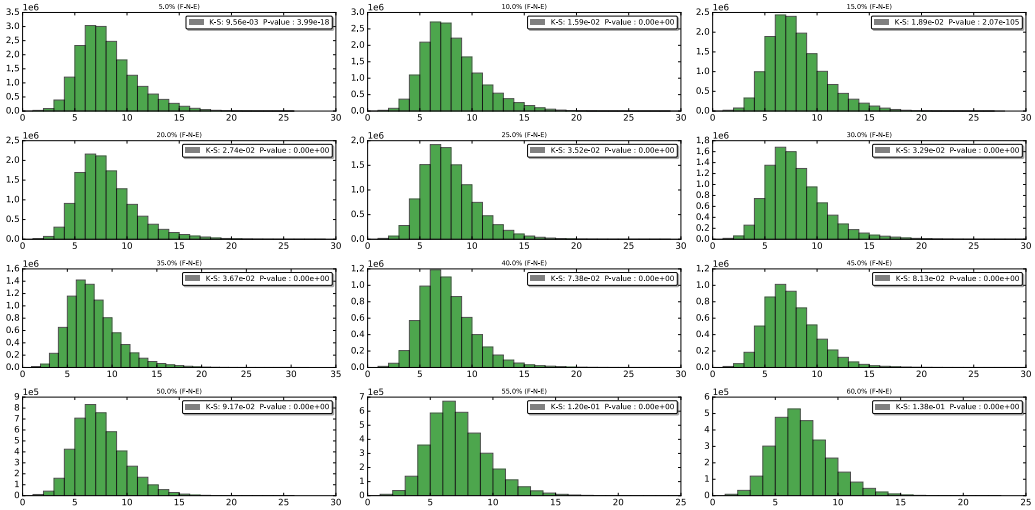
Table 8: Behavior of the Italian giant component at 60% error.

	Nodes	Edges	Nodes Gc	Edges Gc	Density
FNE	4,483	5,007	1,675 (37.3%)	4,320 (86.3%)	0.0031
FNEB	4,483	5,007	2,226 (49.6%)	3,502 (69.9%)	0.0014

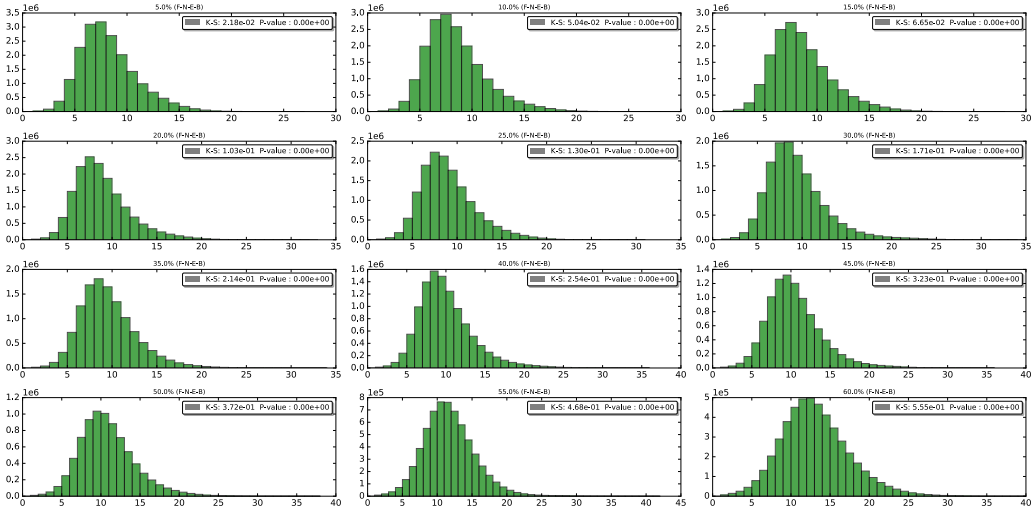
Table 9: Size of the perturbed networks under false negative edges random and biased at 60% of error rate. The percentages reported are relative to the first and second column of the table.

6.1.7 False negative edges - Revenue bias

The result of the assortativity coefficient here is remarkable: we obtain a growth of almost 100%, going from a weak assortativity of 0.30 up to an assortativity of 0.57. The growth process can easily be seen from the plots in Figure 17, where the increasingly growing linearity in the relation between the source degrees and the target degrees. Also, the vertical axis remains surprisingly stable at the same value while the horizontal axis naturally diminishes. Eventually, there seem to be only clear changes in the curve from degrees of 25 on, while in the most left part the assortativity remains almost equal throughout each error rate. Knowing that there is non-assortativity in



(a) Distance distribution for false negative edges at random at each error rate.



(b) Distance distribution for false negative edges biased at each error rate.

Figure 14: Distance distributions for false negative edges random and biased at each error rate.

the revenue of the nodes, since the correlation is equal to 0.1, an explanation of this phenomenon might be found, again, looking in Figure 17. It is plausible that the nodes with degree higher than 25 were those most affected by the removal of links. One can see how the neighbors of the nodes with degree

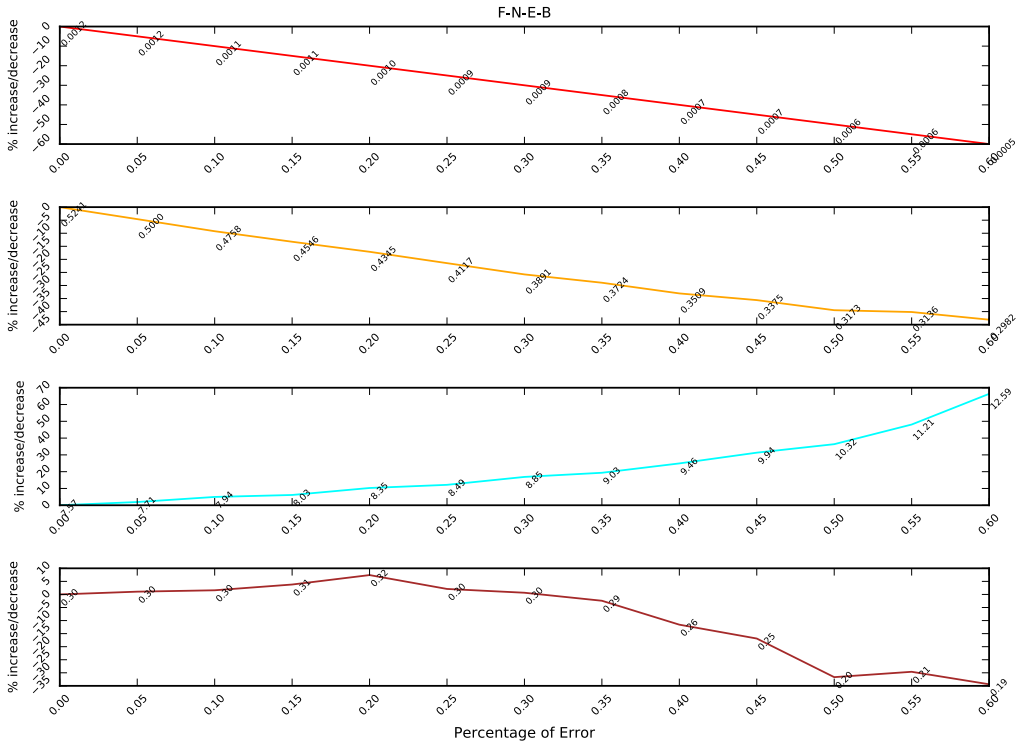


Figure 15: Measure changes under false negative edges biased for each error rate.

around $25 - 35$ fade as we increase the error and how the hubs diminish in size.

There are not noteworthy results regarding the centrality correlations.

6.1.8 False positive nodes - Random

We immediately see how the density remains pretty stable to its original value, as in "false negative nodes at random" — meaning that for graphs large enough the ratio $\frac{m}{n^2}$ with m number of edges and n number of nodes, remains equal — while the assortativity coefficient decreases up to 40%.

Looking at Figure 19 another interesting result is the more than linear decrease of the average distance, which lowered from the initial value of 7.57 to 4.29 at 60% of error. The changes in the average distance can be better understood from the distance distribution plots in Figure 20a, while they

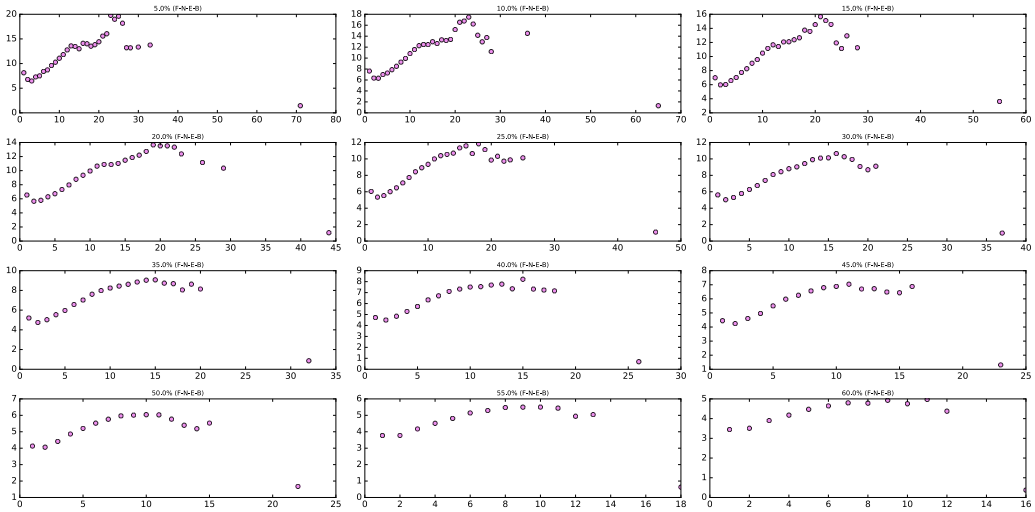


Figure 16: Average neighbor correlation under false negative edges biased towards the degree.

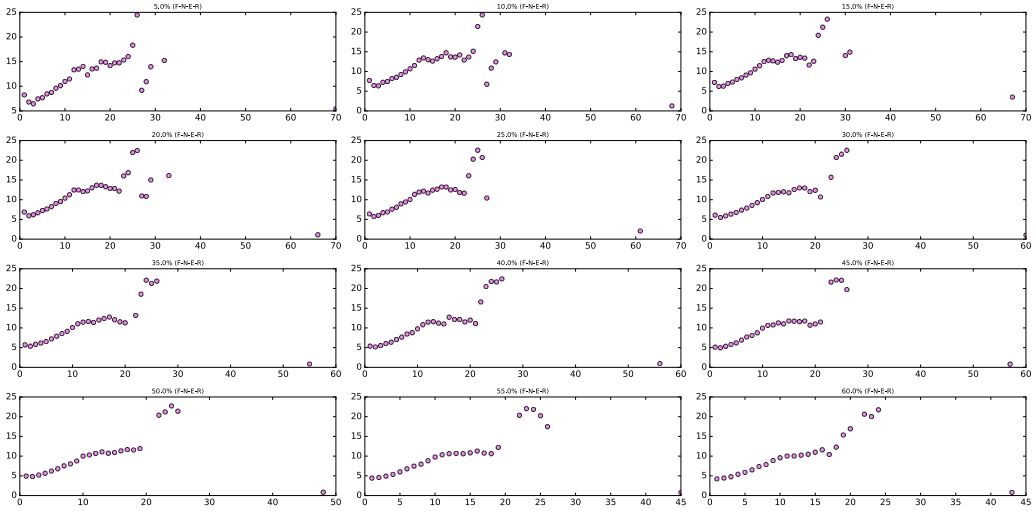


Figure 17: Average neighbor correlation under false negative edges biased towards the revenue.

may be explained by looking at the behavior of the degree distributions in Figure 20b. From the degree distributions we see how the power-law starts to become like a Poisson distribution. The reason for this is the increasing

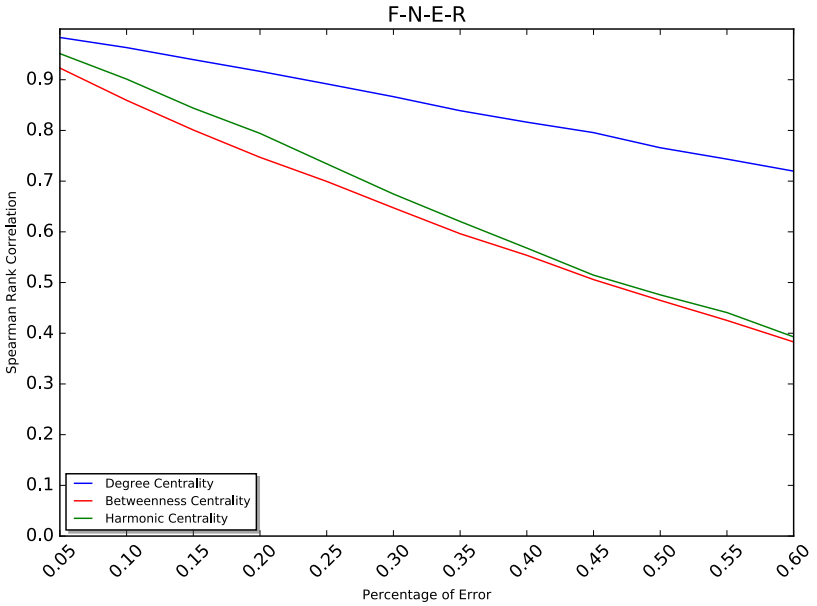


Figure 18: Spearman correlation of degree centrality, betweenness centrality and harmonic centrality for false negative edges biased towards the revenue.

average degree. We recall how for each disconnected node introduced we select a random node and we take its degree (k), which will likely be low. We then connect an edge from the spurious node to other k random nodes (which will also likely have low degree), that will in turn see their degree rise. This procedure eventually helps the average distance to diminish given the more shortest paths that are now possible to follow. On the contrary, adding nodes with average degree equal to 1 would have increased the number of peripheral nodes and consequently the average distance of the network.

For what concerns the Spearman correlation results (see Figure 21a) we see a significant drop of the betweenness and harmonic centrality at 5% of error which decrease from 1 to 0.78 and 0.83 respectively. The degree centrality remains the most stable of the three (reaching a correlation of 0.82 at 60%), while betweenness and harmonic centrality behave very similar with only negligible differences, reaching values of 0.46 and 0.5.

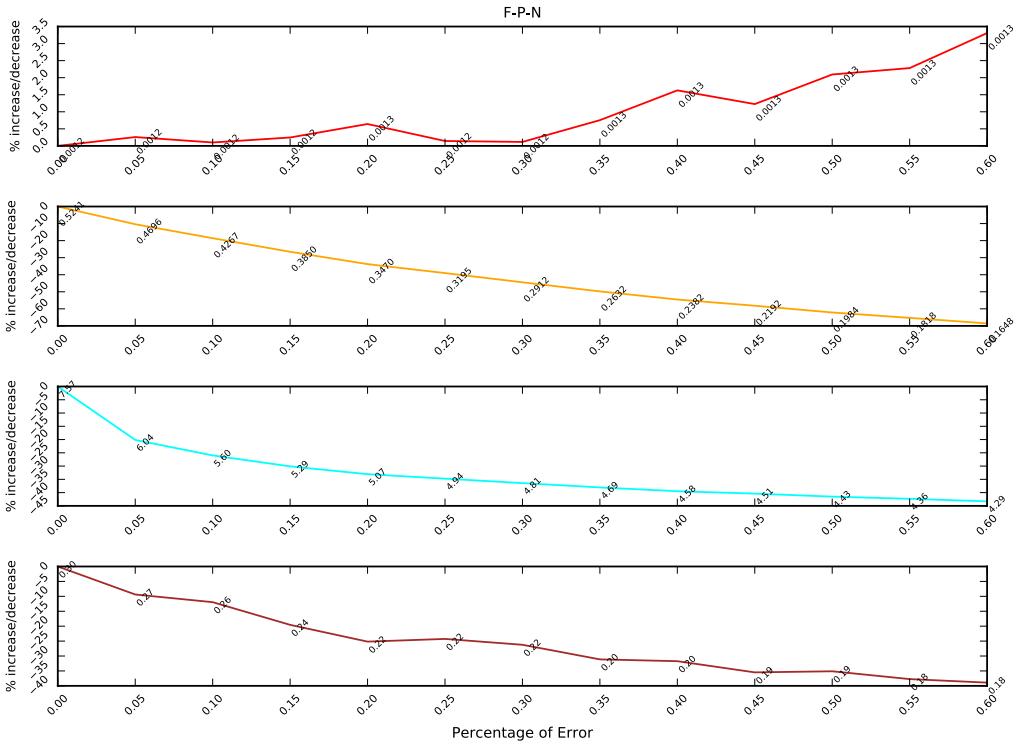
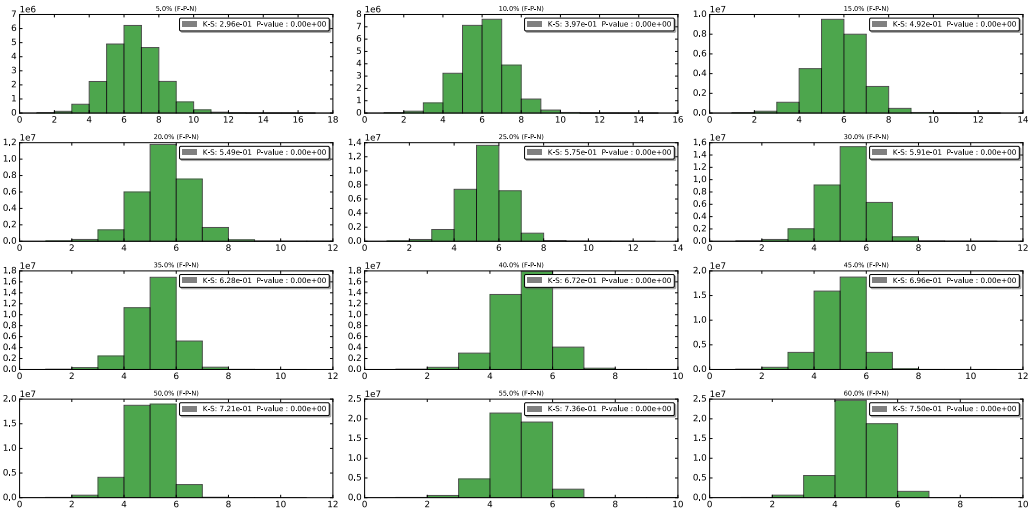


Figure 19: Measure changes under false positive nodes random for each error rate.

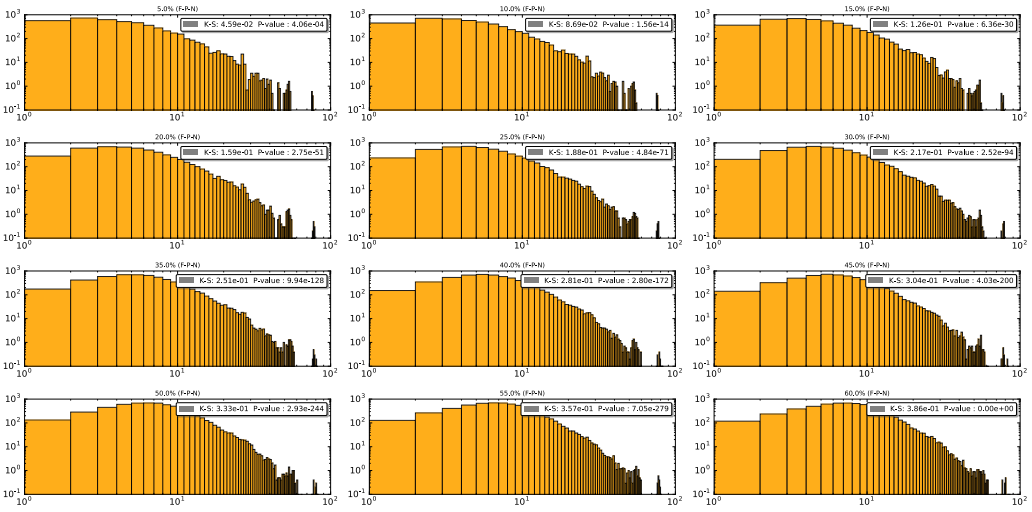
6.1.9 False positive nodes - Degree bias

From the global measures we notice a similarity with Figure 19. The density remains stable (with only negligible changes) throughout all experiments. The clustering coefficient decreases while the average distance decreases from 7.57 to 3.94. We see major differences in the assortativity coefficient: the perturbed graphs H become more non-assortative with respect to the perturbed graphs in the random case.

The results regarding the ranking correlations (see Figure 21a) remain very similar to the ones presented in Figure 21b. The only differences here are the 10% improvement of all correlations with respect to the previous error scenario and the swap in the ranking of betweenness and harmonic centrality. Finally, from Figure 22 we see how the degree distributions do not resemble Poisson distributions anymore.



(a) Distance distribution for false positive nodes random at each error rate.

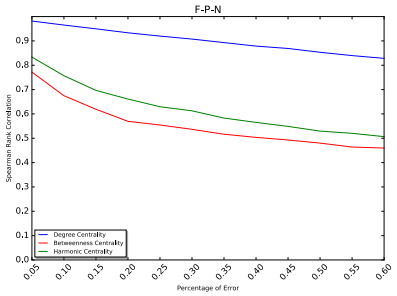


(b) Degree distribution for false positive nodes random at each error rate.

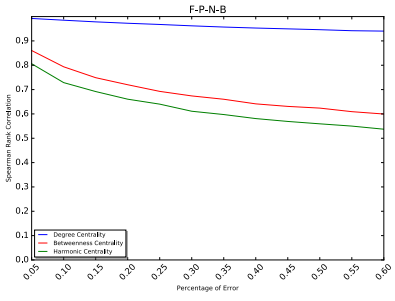
Figure 20: Distance distributions and degree distribution for false positive nodes random at each error rate.

6.1.10 False positive edges - Random

The density increases proportionally to the error rate, as expected. The average distance naturally decreases, while the global clustering coefficient surprisingly drops almost proportionally to the error rate (about -50% at



(a) Spearman correlations for false positive nodes at random.



(b) Spearman correlations for false positive nodes biased.

Figure 21: Spearman correlation of degree centrality, betweenness centrality and harmonic centrality for false positive nodes at random and false positive nodes biased.

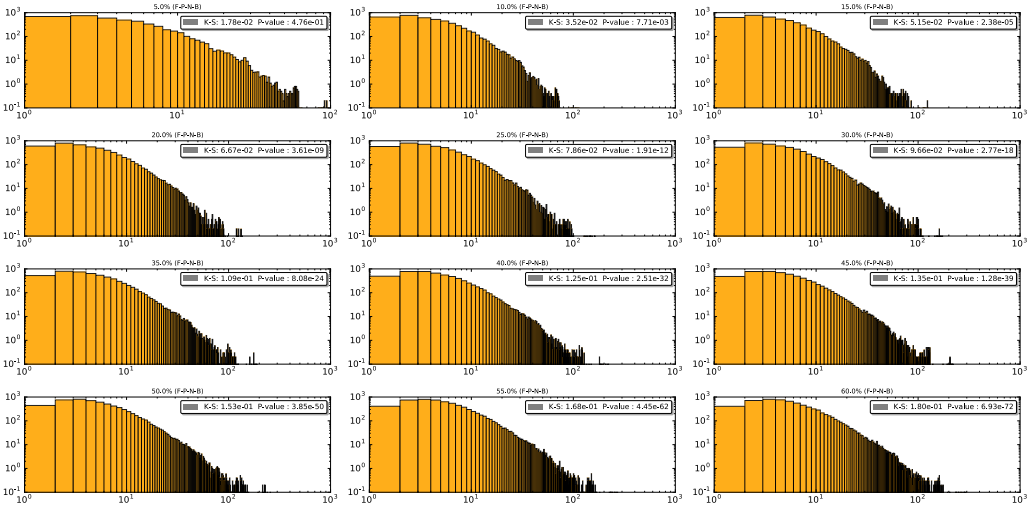


Figure 22: Degree distribution for false positive nodes biased at each error rate.

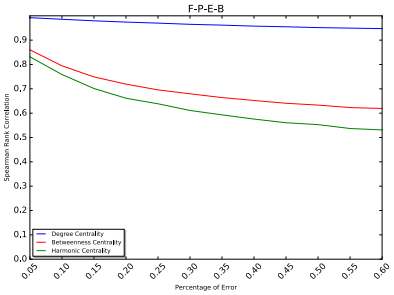
60% of error). If one would expect an increase of the clustering coefficient due to the enclosure of the connected triples that will consequently become triangles, the results suggest that the opposite is actually happening. Namely, selecting two unconnected nodes at random the probability to chose nodes having at least a neighbor in common is lower than the probability to find two completely unrelated nodes.

Eventually, also the degree distribution obtained is pretty much identical to the one we obtain in "false positive nodes at random". A more Poisson-like degree distribution is indeed present.

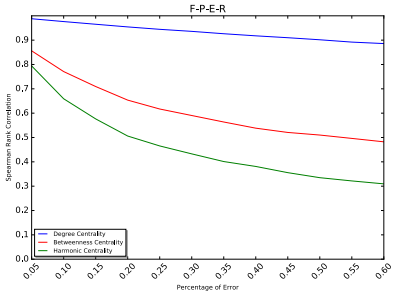
6.1.11 False positive edges - Degree and Revenue biases

The perturbed graphs become non-assortative (-70% and -120% respectively) and the average distance, as well as the clustering coefficient, decrease by $60 - 70\%$.

Regarding the Spearman correlation results in Figure 23a and Figure 23b, degree centrality behaves very similar in both biased cases. Betweenness and harmonic centrality are more stable when the bias is applied towards high degree companies instead of high revenue ones. In particular, in the first case (see Figure 23a) they both behave very similarly (apart from negligible differences), while in the latter (see Figure 23b) harmonic centrality is more unstable than betweenness centrality.



(a) Spearman correlations for false positive edges with bias towards high degree nodes.



(b) Spearman correlations for false positive edges with bias towards high revenue companies.

Figure 23: Spearman correlation of degree centrality, betweenness centrality and harmonic centrality for false positive edges biased towards high degree nodes (a) and high revenue nodes (b).

6.1.12 False aggregation - Random

We forthwith notice how the density increases by a remarkable 500% (see Figure 24). The reason for this is given by the nature of the error scenario: we select two different nodes (one to maintain and one to remove) and we

then connect all the neighbors of the node to remove to the one to keep, avoiding self-loops. At the end we delete all possible parallel edges.

Let us explain the change in density with an example. Let us start from the Italian giant component, which has $4,483$ nodes and $12,517$ edges. Imagine then to aggregate the 60% of the nodes, which will let the number of nodes decrease by 60%, going from $4,483$ to $1,794$. Now, in the best case scenario we will leave the number of edges unchanged at $12,517$. Computing then the density one can easily see how this has grown from 0.0012 to 0.00778, reporting an increase of about 548%. Deleting parallel edges at the end of the process, we generally cut off a few edges, lowering the density from its maximum possible value.

Looking now at the global clustering coefficient changes we see how it decreases proportionally to the error rates. This means that at each error rate we remove the relative percentage of triangles in the network, leaving the numbers of connected triples unchanged. Eventually, the average distance decreases as expected.

6.1.13 False aggregation - Degree bias

The behavior of the global measures in this biased case is very similar to the random one, but the differences end there. Looking at the results for the Spearman correlation reported in Figure 25a and Figure 25b, we see how the ranking of the three errors decrease almost linearly. There is no correlation between the rankings in the original network and in the perturbed network at 60% of error.

6.1.14 False disaggregation - Random

Looking at the results in Figure 26 and Figure 27a, we see how this error scenario is definitely the one that leaves the perturbed network more similar to the original. We register only negligible changes in the clustering coefficient, average distance and assortativity coefficient. Almost perfect Spearman correlations are present.

6.1.15 False disaggregation - Degree bias

Introducing a bias towards high degree companies, the situation does not change much. We see an almost linear decrease in the density from 0.0012 to 0.005, the global clustering coefficient decreases up to -25% at 60% of

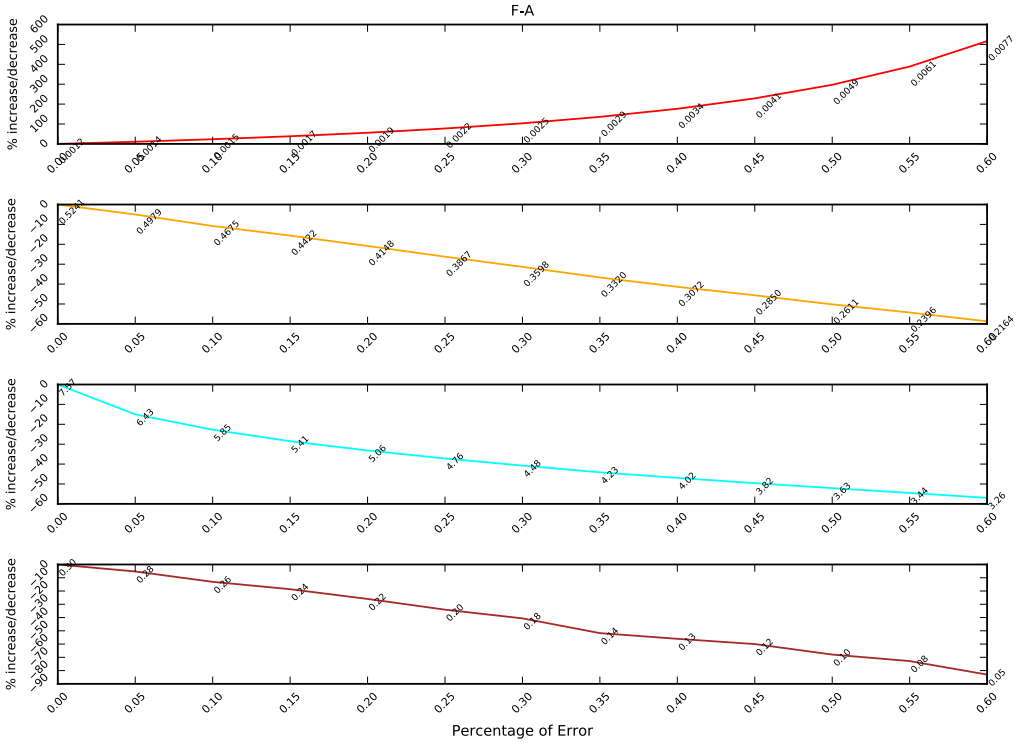
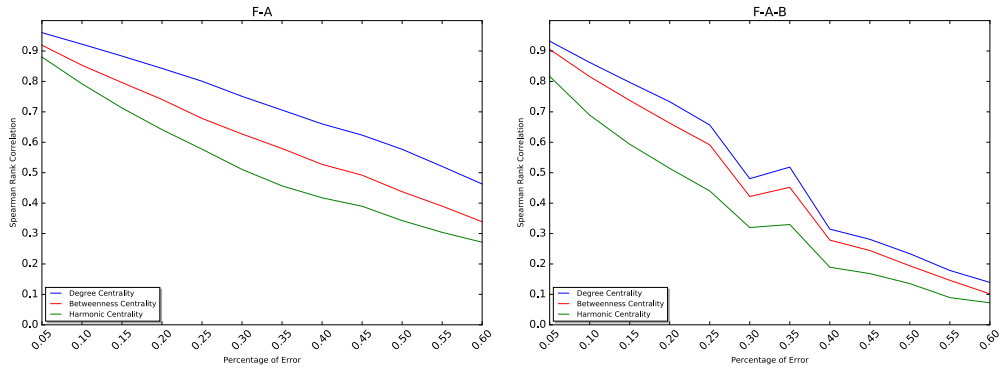


Figure 24: Measure changes under false aggregation at random.

error, the average distance increases up to 12% and the assortativity coefficient remains pretty stable apart from some negligible changes. The same holds for the Spearman correlation results in Figure 27b, which, apart from only negligible differences are identical to the ones obtained in the previous random error scenario.

	Nodes	**Edges**	**Nodes Gc**	**Edges Gc**	**# of components**
FD	7,172	12,517	5,009 (69.84%)	12,294 (98.2%)	199
FDB	7,172	12,517	5,403 (75.33%)	12,156 (97.11%)	152

Table 10: Size of the perturbed networks under false disaggregation random and biased at 60% of error. The percentages reported are relative to the first and second column of the table.



(a) Spearman correlations for false aggregation at random.

(b) Spearman correlations for false aggregation with bias towards high degree nodes

Figure 25: Spearman correlation of degree centrality, betweenness centrality and harmonic centrality for false aggregation at random (a) and biased towards high degree nodes (b).

6.1.16 Community structure results

Here we present the variation of information results computed on the community structure of G and H, as explained in Section 5.4.2.

Looking at the values presented in Table 11 we see how for the majority of the error scenarios the values are distributed following a logarithmic curve. The results are very much consistent with the ones found in Section 6.1, namely: false aggregation disrupts G very aggressively, making G and H look almost completely different. Node addition, edge addition and edge removal act on G in a generally fair way. Node removal and node splitting maintain all the properties of G pretty much throughout all the error rates.

This trend is indeed maintained in Table 11: the variation of information computed under the false aggregation error scenarios at 60% of error present values of 0.69 and 0.78. The two community structures are almost completely different. Despite this, all the other results computed at 60% of error have values in the range $0.13 - 0.47$. For this network, its community structure is generally preserved even under poor data quality conditions.

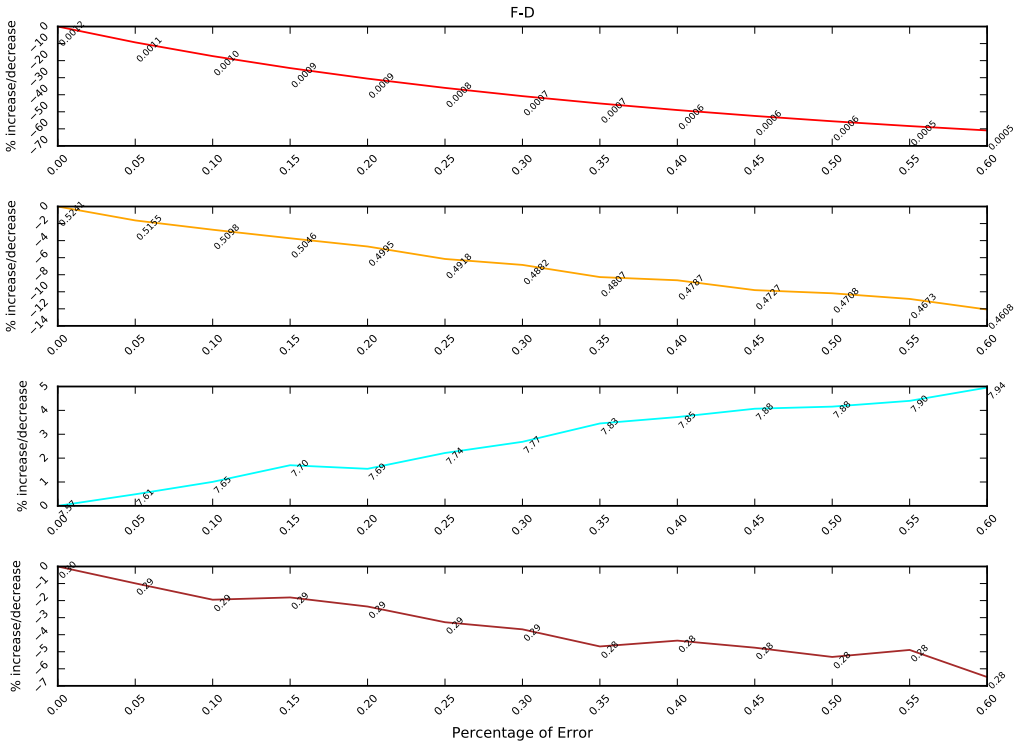
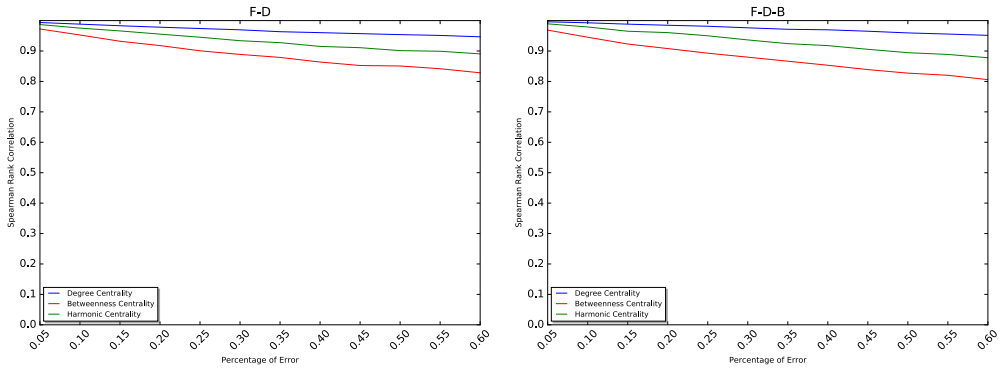


Figure 26: Measure changes under false disaggregation at random.

	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%
FNNR	0.09	0.12	0.14	0.17	0.19	0.21	0.23	0.24	0.26	0.27	0.28	0.30
FPER	0.10	0.16	0.20	0.22	0.24	0.26	0.28	0.30	0.30	0.32	0.33	0.35
FNER	0.09	0.12	0.16	0.19	0.23	0.25	0.29	0.30	0.32	0.34	0.36	0.38
FNN	0.10	0.14	0.17	0.20	0.23	0.24	0.26	0.27	0.29	0.30	0.32	0.32
FPN	0.18	0.25	0.30	0.32	0.36	0.38	0.40	0.43	0.43	0.45	0.46	0.47
FNE	0.10	0.14	0.17	0.21	0.24	0.26	0.29	0.32	0.34	0.35	0.38	0.39
FPE	0.15	0.22	0.27	0.31	0.32	0.35	0.36	0.38	0.39	0.40	0.41	0.42
FA	0.22	0.32	0.38	0.43	0.48	0.52	0.58	0.61	0.63	0.65	0.68	0.69
FD	0.03	0.06	0.07	0.08	0.10	0.10	0.10	0.12	0.12	0.12	0.13	0.13
FNNB	0.07	0.09	0.11	0.12	0.14	0.16	0.17	0.19	0.19	0.20	0.21	0.23
FPNB	0.11	0.17	0.21	0.23	0.25	0.30	0.31	0.32	0.35	0.36	0.38	0.40
FNEB	0.09	0.12	0.14	0.17	0.20	0.22	0.23	0.25	0.27	0.28	0.31	0.33
FPEB	0.10	0.14	0.17	0.19	0.21	0.23	0.26	0.27	0.29	0.29	0.30	0.32
FAB	0.25	0.36	0.43	0.50	0.55	0.63	0.64	0.70	0.72	0.74	0.76	0.78
FDB	0.04	0.07	0.08	0.09	0.10	0.11	0.12	0.12	0.13	0.14	0.14	0.16

Table 11: Variation of information results normalized to $[0, 1]$ for communities in the Italian network's giant component.



(a) Spearman correlations for false disaggregation at random.

(b) Spearman correlations for false disaggregation with bias towards high degree nodes.

Figure 27: Spearman correlation of degree centrality, betweenness centrality and harmonic centrality for false disaggregation at random and biased towards high degree nodes.

6.2 Country resilience

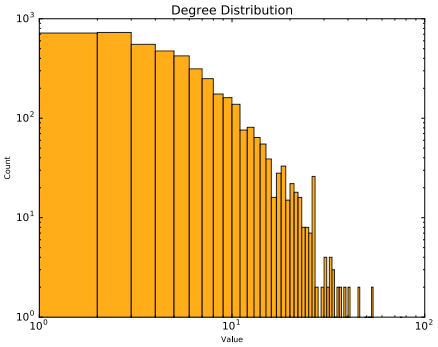
Where as for the Italian giant component we presented a detailed analysis of the results obtained, from this point on we will abstract away. Here we propose an easily and effective way to visualize and understand how resilient a certain network is under the data quality artifacts. Indeed, we will present the results of the Italian, Danish, UK, Scandinavian, Dutch and Spanish corporate network's giant components by means of what we called a *resilience matrix*. Before entering in the details of such a matrix, in Table 12, Figure 28 and Figure 29 we present some of the topological features of the networks we analyze. Eventually, the names of the networks have been encoded by means of the *ISO 3166-1 alpha-2* codes as follows:

- IT: Italian network
- DK: Danish network
- UK: United Kingdom network
- NL: Dutch network
- ES: Spanish networks

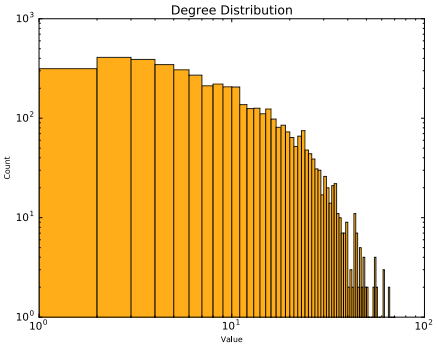
We will refer to the Scandinavian network — consisting of Denmark, Norway, Sweden, Finland and Iceland together — with the symbol "SCA".

	IT	**DK**	**UK**	**NL**	**ES**	**SCA**
Nodes	4,483	4,517	32,962	6,083	11,102	25,765
Edges	12,517	23,381	366,381	50,107	87,907	146,166
Density	0.0012	0.0022	0.00067	0.0027	0.0014	0.0004
Global c.c.	0.524	0.089	0.14	0.13	0.14	0.109
Avg. Dist	7.56	5.61	6.62	7.61	6.29	6.66
Assort. coeff.	0.29	0.40	0.87	0.78	0.87	0.63

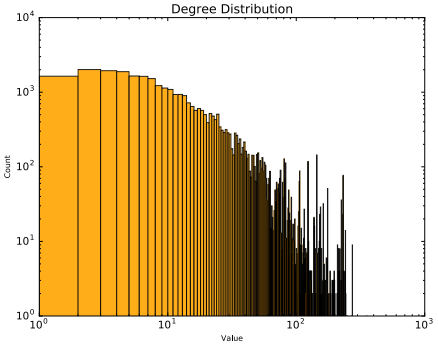
Table 12: Topological features of the Italian, Danish, UK, Dutch, Spanish and Scandinavian networks.



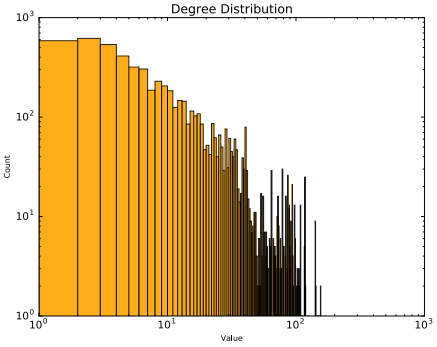
(a) Degree distribution of the IT network



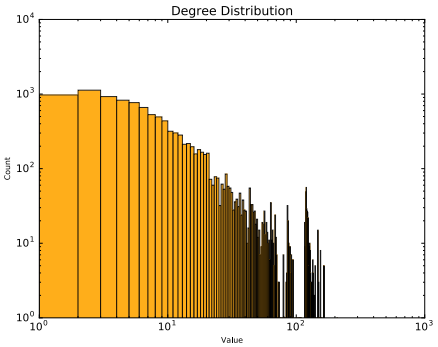
(b) Degree distribution of the DK network



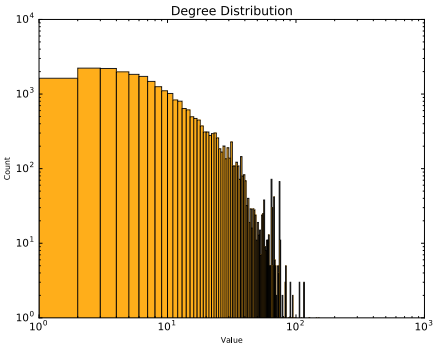
(c) Degree distribution of the UK network



(d) Degree distribution of the NL network

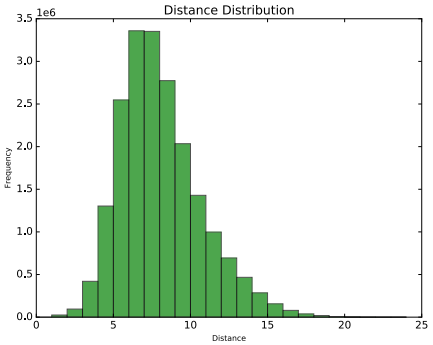


(e) Degree distribution of the ES network

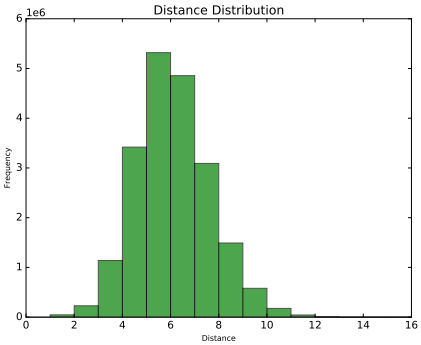


(f) Degree distribution of the SCA network

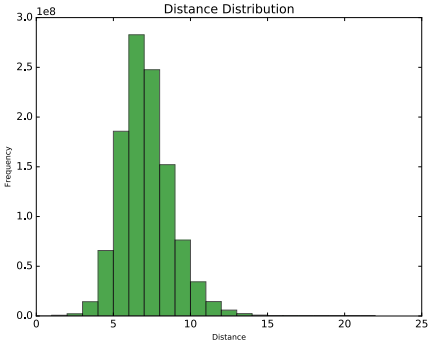
Figure 28: Degree distributions of the network's giant components analyzed.



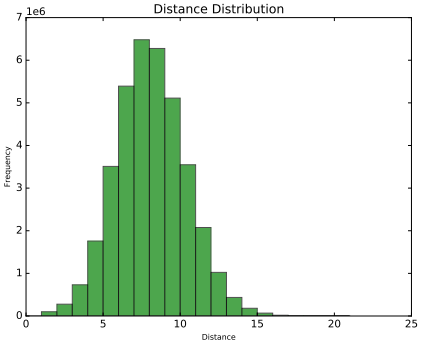
(a) Distance distribution of the IT network



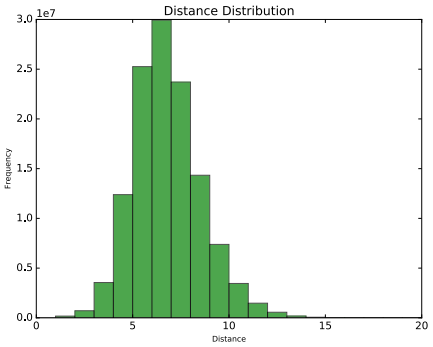
(b) Distance distribution of the DK network



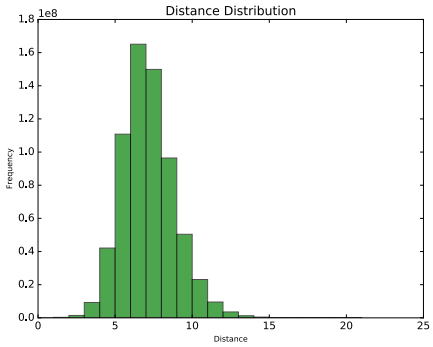
(c) Distance distribution of the UK network



(d) Distance distribution of the NL network



(e) Distance distribution of the ES network



(f) Distance distribution of the SCA network

Figure 29: Distance distributions of the networks analyzed.

To understand whether and how these features play a role in the resilience, to study what error scenarios are more (or less) aggressive and to finally understand whether the networks still maintain their integrity under poor data quality, in Table 13 and Table 14 we present the *resilience matrix* of the networks.

We define the *resilience matrix* as a matrix describing how "different" a perturbed network H is from its original version G when 60% of error is introduced.

The concept of "difference" is given by the measures used before: Spearman correlation values for the degree, betweenness and harmonic centrality, KS test on the distributions, variation of information on the community structures and finally the density, global clustering coefficient, average distance and assortativity coefficient percentages increase/decrease. For sake of clarity and simplicity we represent their respective values using the following notation:

- **Spearman correlation**
  - ● $\rightarrow (0.66; 1]$
  - ● $\rightarrow (0.33; 0.66]$
  - ● $\rightarrow [0; 0.33]$
- **VI & KS 2 sample test**
  - ● $\rightarrow [0; 0.33]$
  - ● $\rightarrow (0.33; 0.66]$
  - ● $\rightarrow (0.66; 1]$
- **Percentages increase**
  - ▲ $\rightarrow$ Increase in $[0\% - 33\%]$
  - ▲ $\rightarrow$ Increase in $(33\% - 66\%]$
  - ▲ $\rightarrow$ Increase in $(66\% - 100\%]$
  - ▲ $\rightarrow$ Increase more than $+100\%$
- **Percentages decrease**
  - ▼ $\rightarrow$ Decrease in $[0\% - 33\%]$

▼ → Decrease in $(33\% - 66\%]$

▼ → Decrease in $(66\% - 100\%]$

▼ → Decrease more than 100%

	Network	Deg. Centr	Betw. Centr	Harm. Centr	VI	KS Deg	KS Dist	Density (%)	Global c.c. (%)	Avg. Dist. (%)	Assort. coeff. (%)
FNNR	IT	●	●	●	●	●	●	▲	▼	▼	▼
	DK	●	●	●	●	●	●	▼	▲	▲	▲
	UK	●	●	●	●	●	●	▲	▲	▲	▲
FPER	IT	●	●	●	●	●	●	▲	▼	▼	▼
	DK	●	●	●	●	●	●	▲	▼	▼	▼
	UK	●	●	●	●	●	●	▲	▼	▼	▼
FNER	IT	●	●	●	●	●	●	▼	▲	▲	▲
	DK	●	●	●	●	●	●	▼	▼	▼	▼
	UK	●	●	●	●	●	●	▼	▼	▲	▼
FNN	IT	●	●	●	●	●	●	▼	▲	▼	▲
	DK	●	●	●	●	●	●	▼	▼	▲	▲
	UK	●	●	●	●	●	●	▼	▼	▲	▼
FPN	IT	●	●	●	●	●	●	▲	▼	▼	▼
	DK	●	●	●	●	●	●	▲	▼	▼	▼
	UK	●	●	●	●	●	●	▲	▼	▼	▼
FNE	IT	●	●	●	●	●	●	▼	▼	▼	▲
	DK	●	●	●	●	●	●	▼	▼	▼	▼
	UK	●	●	●	●	●	●	▼	▼	▲	▲
FPE	IT	●	●	●	●	●	●	▲	▼	▼	▼
	DK	●	●	●	●	●	●	▲	▼	▼	▼
	UK	●	●	●	●	●	●	▲	▼	▼	▼
FA	IT	●	●	●	●	●	●	▲	▼	▼	▼
	DK	●	●	●	●	●	●	▲	▼	▼	▼
	UK	●	●	●	●	●	●	▲	▼	▼	▼
FD	IT	●	●	●	●	●	●	▼	▼	▲	▼
	DK	●	●	●	●	●	●	▼	▼	▲	▼
	UK	●	●	●	●	●	●	▼	▼	▲	▼
FNNB	IT	●	●	●	●	●	●	▲	▲	▼	▲
	DK	●	●	●	●	●	●	▲	▲	▼	▲
	UK	●	●	●	●	●	●	▲	▲	▼	▲
FPNB	IT	●	●	●	●	●	●	▲	▼	▼	▼
	DK	●	●	●	●	●	●	▲	▼	▼	▼
	UK	●	●	●	●	●	●	▲	▼	▼	▼
FNEB	IT	●	●	●	●	●	●	▼	▼	▲	▼
	DK	●	●	●	●	●	●	▼	▼	▲	▼
	UK	●	●	●	●	●	●	▼	▼	▲	▼
FPEB	IT	●	●	●	●	●	●	▲	▼	▼	▼
	DK	●	●	●	●	●	●	▲	▼	▼	▼
	UK	●	●	●	●	●	●	▲	▼	▼	▼
FAB	IT	●	●	●	●	●	●	▲	▼	▼	▼
	DK	●	●	●	●	●	●	▲	▼	▼	▼
	UK	●	●	●	●	●	●	▲	▼	▼	▼
FDB	IT	●	●	●	●	●	●	▼	▼	▲	▼
	DK	●	●	●	●	●	●	▼	▼	▲	▼
	UK	●	●	●	●	●	●	▼	▼	▲	▼

Table 13: Resilience matrix for the Italian, Danish and UK network's giant components.

	Network	Deg. Centr	Betw. Centr	Harm. Centr	Var. Info.	KS Deg	KS Dist	Density (%)	Global c.c. (%)	Avg. Dist. (%)	Assort. coeff. (%)
FNNR	NL	●	●	●	●	●	●	▲	▲	▲	▼
	ES	●	●	●	●	●	●	▲	▲	▲	▼
	SCA	●	●	●	●	●	●	▲	▲	▲	▲
FPER	NL	●	●	●	●	●	●	▲	▼	▼	▼
	ES	●	●	●	●	●	●	▲	▼	▼	▼
	SCA	●	●	●	●	●	●	▲	▼	▼	▼
FNER	NL	●	●	●	●	●	●	▼	▼	▲	▼
	ES	●	●	●	●	●	●	▼	▼	▼	▼
	SCA	●	●	●	●	●	●	▼	▼	▼	▲
FNN	NL	●	●	●	●	●	●	▼	▼	▲	▼
	ES	●	●	●	●	●	●	▼	▼	▲	▼
	SCA	●	●	●	●	●	●	▼	▼	▲	▲
FPN	NL	●	●	●	●	●	●	▲	▼	▼	▼
	ES	●	●	●	●	●	●	▲	▼	▼	▼
	SCA	●	●	●	●	●	●	▲	▼	▼	▼
FNE	NL	●	●	●	●	●	●	▼	▼	▼	▼
	ES	●	●	●	●	●	●	▼	▼	▼	▼
	SCA	●	●	●	●	●	●	▼	▼	▼	▼
FPE	NL	●	●	●	●	●	●	▲	▼	▼	▼
	ES	●	●	●	●	●	●	▲	▼	▼	▼
	SCA	●	●	●	●	●	●	▲	▼	▼	▼
FA	NL	●	●	●	●	●	●	▲	▼	▼	▼
	ES	●	●	●	●	●	●	▲	▼	▼	▼
	SCA	●	●	●	●	●	●	▲	▼	▼	▼
FD	NL	●	●	●	●	●	●	▼	▼	▲	▼
	ES	●	●	●	●	●	●	▼	▼	▲	▼
	SCA	●	●	●	●	●	●	▼	▼	▲	▼
FNNB	NL	●	●	●	●	●	●	▲	▲	▲	▼
	ES	●	●	●	●	●	●	▲	▲	▼	▼
	SCA	●	●	●	●	●	●	▲	▲	▼	▲
FPNB	NL	●	●	●	●	●	●	▲	▼	▼	▼
	ES	●	●	●	●	●	●	▲	▼	▼	▼
	SCA	●	●	●	●	●	●	▲	▼	▼	▼
FNEB	NL	●	●	●	●	●	●	▼	▼	▲	▼
	ES	●	●	●	●	●	●	▼	▼	▲	▼
	SCA	●	●	●	●	●	●	▼	▼	▲	▼
FPEB	NL	●	●	●	●	●	●	▲	▼	▼	▼
	ES	●	●	●	●	●	●	▲	▼	▼	▼
	SCA	●	●	●	●	●	●	▲	▼	▼	▼
FAB	NL	●	●	●	●	●	●	▲	▼	▼	▼
	ES	●	●	●	●	●	●	▲	▼	▼	▼
	SCA	●	●	●	●	●	●	▲	▼	▼	▼
FDB	NL	●	●	●	●	●	●	▼	▼	▲	▼
	ES	●	●	●	●	●	●	▼	▼	▲	▼
	SCA	●	●	●	●	●	●	▼	▼	▲	▼

Table 14: Resilience matrix for the Dutch, Spanish and Scandinavian network's giant components.

Even though the set of networks used is pretty heterogeneous (in terms of network's properties), in Table 13 and Table 14 a strong heterogeneity in the results is not present. All of them reacted mostly the same, given an error scenario. Namely, given a certain error scenario is very difficult to see different networks reacting in opposite ways.

In order to better understand which scenarios disrupt the networks the most, from the latter two tables we decide to generate Table 15 obtained by summing the results of every network under each artifact.

	Deg. Centr	Betw. Centr	Harm. Centr	Var. Info.	KS Deg	KS Dist	Density (%)	Global c.c. (%)	Avg. Dist. (%)	Assort. coeff. (%)
FNNR	6●	6●	3●3●	4●2●	4●2●	3●3●	4▲1▼1▲	5▲1▼	3▲2▲1▼	3▲3▼
FNN	6●	6●	3●3●	3●3●	3●3●	3●3●	6▼	5▼1▲	4▲1▼1▲	3▲3▲
FNNB	6●	6●	6●	6●	6●	6●	6▲	6▲	5▼1▲	3▲2▼1▲
FPER	6●	4●2●	3●2●1●	5●1●	5●1●	6●	6▲	5▼1▼	6▼	5▼1▼
FPE	6●	5●1●	4●2●	5●1●	6●	5●1●	6▲	5▼1▼	6▼	6▼
FPEB	6●	4●1●1●	4●1●1●	6●	6●	6●	6▲	3▼3▼	6▼	4▼2▼
FNER	6●	3●3●	5●1●	5●1●	3●3●	6●	6▼	5▼1▲	3▲3▼	4▼1▲1▲
FNE	6●	3●3●	4●1●1●	5●1●	3●3●	6●	6▼	6▼	5▼1▲	4▼2▲
FNEB	6●	6●	4●2●	5●1●	6●	4●2●	6▼	6▼	4▲1▲1▲	3▼3▼
FPN	6●	5●1●	6●	5●1●	6●	6●	6▲	3▼3▼	6▼	6▼
FPNB	6●	5●1●	5●1●	4●2●	6●	6●	6▲	6▼	6▼	4▼2▼
FA	6●	4●2●	4●2●	4●2●	3●3●	5●1●	6▲	6▼	6▼	4▼2▼
FAB	6●	6●	6●	5●1●	5●1●	6●	6▲	6▼	6▼	5▼1▼
FD	6●	6●	6●	6●	6●	6●	5▼1▼	5▼1▼	6▲	6▼
FDB	6●	6●	6●	6●	6●	6●	5▼1▼	6▼	6▲	6▼

Table 15: Resilience matrix visualization obtained by clustering the results of all the six networks together.

From Table 15 some trends seem to be present. In particular, all the "false negative nodes" artifacts at 60% only bear minor changes to H: the difference between G and H is very subtle. The same happens under the "false disaggregation" artifacts, where the networks maintain almost all their properties perfectly intact.

Biased artifacts are a bit less aggressive on G than the revenue biased and the random ones, even if the differences are very minor.

"False positive edges", "false negative edges" and "false positive nodes" are pretty much identically aggressive. Even though the way in which they disrupt the networks is diverse, they all do not preserve the features of the original graph G as much as the aforementioned two. Notice that the density changes in "false positive edges", "false negative edges" are trivial: removing 60% of the edges, the density decreases of 60%.

The community structure, as well as most of the other network measures, under "false positive edges", "false negative edges" and "false positive nodes", might look slightly different from the original one. Under the three afore-

	Deg. Centr	Betw. Centr	Harm. Centr	Var. Info.	KS Deg	KS Dist	Density (%)	Global c.c. (%)	Avg. Dist. (%)	Assort. coeff. (%)
Total	78●	37●	40●	43●	48●	42●	28▼ 19▲	35 ▼ 29 ▼	42▼ 27▲	39▼ 22▼
	6●	31●	35●	40●	36●	40●	18▲ 16▲	13 ▲ 13 ▼	15▼ 5▲ 1▲	12▲ 11▼
	6●	22●	15●	7●	6●	8●	8▼ 1▼			4▲ 1▲ 1▼

Table 16: Number of green, orange, red and black dots or triangles per measure.

mentioned scenarios the variation of information is most of the times in the range of $(0.33, 0.66]$. The degree biased case, instead, at least for "false positive edges" and "false negative edges", preserves the community structure of G almost perfectly.

Our results finally suggest that "false aggregation at random" alters G more than other error scenarios but still much less than its biased version, which makes the original network G almost unrecognizable.

Focusing now on the resilience of network measures, looking at Table 16 we see how degree centrality is far more resistant than betweenness and harmonic centrality. This results is based on the fact that measures like betweenness and harmonic, for how they are defined, are more sensitive to changes. In particular, whereas degree centrality of a node will change if neighbors will be removed or added, his betweenness and harmonic centrality will suffers from less local changes. The latter are pretty dependent on the error scenario: none of the two is significantly more resistant than the other.

6.3 Discussion

We observe less robustness on "false negative edges" scenarios with respect to other "false negative nodes". As explained by Wang et al. (2012) removing nodes at random implies removing low-degree nodes while removing edges at random one removes edges belonging to high-degree nodes. Knowing that these latter usually represent the core of the structure of the network, decreasing their degree can significantly compromise the whole system.

Where both Wang et al. (2012) and Borgatti et al. (2006) suggest a similar robustness of the centrality measures to the artifacts, our results show that this is not the case. Where betweenness and harmonic react almost completely equally, degree centrality is much more robust. Whereas degree centrality only suffers from local changes, betweenness and harmonic centrality of a node may be affected even if very distant changes in the network are

reported.

The same happens to most of the global measures. Their behavior is in some sense predictable. Looking at Table 15, we see how the density, global clustering coefficient and average distance increase (or decrease) similarly, given the same artifact. Knowing the error scenario, thus, it should be possible to *a-priori* predict their range of increase/decrease. Eventually, notice how some of the density changes are trivial: removing 60% of edges, for instance, always lowers the density of 60%. An increase of 60% is instead expected when 60% of edges is added.

Where in Wang et al. (2012), "false aggregation" the results were between "false negative nodes" and "false positive edges" results, here is not the case. Even though "false aggregation" involves removing a node A and attaching A's edges to another node B, its results are worse than "false negative nodes" and "false positive edges". We attribute this behavior to the combined effect of both artifacts. In the long term, they tend to line up the degree distribution replacing medium-low (or medium-high in the biased case) degree nodes with high degree ones.

"False disaggregation", which involves removing a certain number of edges from A to attach them to a new node B, happens to be the least disruptive of all. The whole structure of the graph is maintained, as supported by the more global measures in Table 15. This was not the case in Wang et al. (2012), where there results were between "false negative edges" and "false positive nodes" results.

6.4 Advice for corporate networks researchers

Studying corporate board networks involves making decisions. Depending on the kind of conclusions one wants to derive from the network, both the selection of the dataset and the network's measures are fundamental. Let us imagine a scenario in which more corporate datasets are available and each of them has different data quality issues. Now, knowing how networks react to a certain error scenarios gives the possibility to help make the decision of which dataset and measures is convenient to use. In particular:

- When companies that have split into two are still present as a single company ("false aggregation"), the analysis may be not reliable;
- *A-priori* avoid betweenness and harmonic centrality If there are suspects of poor data quality. Whether possible, focus on the degree in-

stead, which is more robust;

- Community analysis is barely influenced by most of the artifacts. Even if the data is imperfect, in most of the cases the communities remain reliable.

- Consider the observed global measures in an interval. For instance, analyzing the density one can simply increase (or decrease) its observed value by the percentages reported in Table 15, so to have a more precise measurement. The same holds for global clustering coefficient and average distance. Let us suppose to study the global clustering coefficient c in a network generated from a dataset having missing companies. To a more precise measurement of c, which takes into account the missing companies, from the *resilience matrix* in Table 15, one can derive that considering c in an interval $[-33\%, c]$ is safer.

Overall, our results suggest that even if corporate data is imperfect, the approximation between real and observed values is high enough to guarantee significant studies. We consider corporate board interlock networks, or at least the ones considered in this work, resilient enough to still be studied under most data quality issues.

7 Conclusion

In this thesis we extended previous work on the impact of imperfect data in network analysis. We concentrated our efforts on corporate board interlock networks. We tested the robustness of six networks and the changes in their global metrics, division into communities, distributions and centrality measures under fifteen artifacts. We proposed an easy and effective way to describes how different a perturbed network H is from its original version G, which we called *resilience matrix*. From this matrix we observed how community analysis is barely influence by the artifacts and how degree centrality is more robust than betweenness and harmonic. Furthermore, we have shown how the matrix help corporate network scientists to *a priori* determine the range of increase/decrease of most of the global measures considered in this work. We concluded this work suggesting that the resilience of corporate board interlock networks is high enough to still study them even if the data quality artifacts are present.

7.1 Future work

Progress can be made analyzing the effect of mixed artifacts on networks. Studying the robustness of several networks under "false negative nodes" combined with "false aggregation", for instance, might be a further step towards more realistic artifacts. Finally, in the particular case of corporate board interlock networks, given their weighted nature (discussed in Section 4), understanding how weighted networks react to imperfect data may be an important and interesting future task.

References

Arnold, T. B. and Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2):34–39.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceeding of International AAAI Conference on Weblogs and Social Media*.

Battiston, S. and Catanzaro, M. (2004). Statistical properties of corporate board and director networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):345–352.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Borgatti, S. P., Carley, K. M., and Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28(2):124–136.

Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188.

Burt, R. S. (2009). *Structural holes: The social structure of competition*. Harvard University Press.

Caldarelli, G. and Catanzaro, M. (2004). The corporate boards networks. *Physica A: Statistical Mechanics and its Applications*, 338(1):98–106.

Chu, J. S. and Davis, G. F. (2011). Who killed the inner circle? The breakdown of the American corporate elite network, 1999–2009.

Chu, J. S. and Davis, G. F. (2015). Who killed the inner circle? The decline of the American corporate interlock network. *Ross School of Business Paper*.

Chung, F. and Lu, L. (2002). Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145.

Costenbader, E. and Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307.

Croci, E. and Grassi, R. (2014). The economic effect of interlocking directorates in italy: new evidence using centrality measures. *Computational and Mathematical Organization Theory*, 20(1):89–112.

Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae (Debrecen)*, 6:290–297.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

Grassi, R. (2010). Vertex centrality as a measure of information flow in Italian corporate board networks. *Physica A: Statistical Mechanics and its Applications*, 389(12):2455–2464.

Heemskerk, E. M., Fennema, M., and Carroll, W. K. (2016). The global corporate elite after the financial crisis: evidence from the transnational network of interlocking directorates. *Global Networks*, 16(1):68–88.

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

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107.

Kim, M. and Leskovec, J. (2011). The network completion problem: Inferring missing nodes and edges in networks. In *Proceedings of SDM*, pages 47–58. SIAM.

Langville, A. N. and Meyer, C. D. (2012). *Who's # 1?: The science of rating and ranking*. Princeton University Press.

Leskovec, J. and Horvitz, E. (2007). Worldwide buzz: Planetary-scale views on an instant-messaging network. Technical Report MSR-TR-2006-186, Microsoft Research.

Meilă, M. (2007). Comparing clusterings — An information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Mizruchi, M. S. (1996). What do interlocks do? An analysis, critique, and assessment of research on interlocking directorates. *Annual Review of Sociology*, pages 271–298.

Newman, M. E. (2003a). Mixing patterns in networks. *Physical Review E*, 67(2):026126.

Newman, M. E. (2003b). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

Orbis — Bureau van Dijk (2016). Orbis website. `https://orbis.bvdinfo.com/`. [Online; accessed 12-Oct-2016].

Peixoto, T. P. (2014). The graph-tool Python library.

Piccardi, C., Calatroni, L., and Bertoni, F. (2010). Communities in Italian corporate networks. *Physica A: Statistical Mechanics and its Applications*, 389(22):5247–5258.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4):211–218.

Rinaldi, A. and Vasta, M. (2014). Persistent and stubborn. The state in the Italian capitalism: 1913—2001. In *The power of corporate networks: A comparative and historical perspective*, pages 169–188. Routledge Press London.

Rochat, Y. (2009). Closeness centrality extended to unconnected graphs: The harmonic centrality index. In *Proceedings of the 6th Conference of Applications of Social Network Analysis*.

Takes, F. W. and Heemskerk, E. M. (2016). Centrality in the Global Network of Corporate Control. *Social Network Analysis and Mining*, 6(1):article 97, 2016.

Van der Hofstad, R. (2016). Random Graphs and Complex Networks. Vol. I. `http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf`. In preparation.

Vitali, S. and Battiston, S. (2014). The community structure of the global corporate network. *PLoS ONE*, 9(8):1–13.

Wang, D. J., Shi, X., McFarland, D. A., and Leskovec, J. (2012). Measurement error in network data: A re-classification. *Social Networks*, 34(4):396–409.

Watts, D. and Strogatz, S. (1998). Collective dynamics of small-world networks. *Nature*, 393:440–442.

Windolf, P. (2014). The corporate network in Germany, 1896–2010. In *The power of corporate networks: A comparative and historical perspective*, pages 66–85. Routledge Press London.

Zafarani, R., Abbasi, M. A., and Liu, H. (2014). *Social media mining: An introduction*. Cambridge University Press.