Potential Biases in Network Reconstruction Methods Not Maximizing Entropy

Rachkov, A.; Pijpers, F.P.; Garlaschelli, D.

Publication date
2021
Document Version
Final published version
License
Unspecified

Discussion Paper

Potential Biases in Network Reconstruction Methods Not Maximizing Entropy

A. Rachkov, F.P. Pijpers, D. Garlaschelli

Mar 2021

Contents

1 Introduction

In recent years the study of complexity and complex networks has been shown to be highly valuable, not only from a scientific or academic point of view, but also from a societal one. For instance, the 2007-2009 financial crisis instigated a strong desire to successfully evaluate the systemic risk, or vulnerability, of a financial system with the hopes to guide regulatory frameworks such that they reduce the danger of future crises. To do so, it is necessary to know the structure of the financial system in question. The primary problem with this is that such information is not normally publicly available. Indeed, privacy and confidentiality rules and regulations frequently hinder the complete access to the required information on interactions between financial agents. For this reason, many methods within the field of complexity theory have focused on reproducing systems from partial information. It is important to realize, however, that merely having such methods is not sufficient. In practice, it is also vital that one understands the behavior and limitations posed by such methods and how they fare with respect to each other. This is precisely the focal point of the current paper.

The Dutch national statistical institute , Statistics Netherlands (SN), has in recent years expressed its own interest in reproducing the structure of economic and financial networks from limited information and has to that end developed its own method to do so, cf. (Hooijmaaijers and Buiten, 2019). Before routinely using any method for producing output for national statistics, however, it is necessary to comprehend how successful such methods are in achieving the desired goal. Thus, to assess the behavior of the method described in Hooijmaaijers and Buiten (2019), its performance is compared to that of methods developed within the framework known for producing the least biased system structures with respect to the available information, known as the maximum entropy framework.

1.1 Relevance and Prior Work

The issue of incomplete data has continuously posed challenges for the production of official statistics. The use of complexity theory in official statistics is thus no exception, where this problem often manifests itself as missing information on existing nodes or links within a network. Unfortunately, this leads to the undesirable situation where researchers are left with a sample constituting a non-representative portion of the population network of interest (Hsieh et al. (2018)). The outcome of such situations is normally biased and therefore unsuitable for publication as official statistics (Bliss et al. (2014); Huisman (2009); Kim and Leskovec (2011); Smith et al. (2017)).

Key studies looking into the effect of missing data in the estimation of global network properties, have found that an increase in missing network data inevitably leads to an increase in estimation bias (Smith et al. (2017); Smith and Moody (2013); Borgatti et al. (2006)). Methods do exist that circumvent the unwanted bias in estimating network properties which allow one to scale estimates from partially observed networks to the target population network (e.g. (Bliss et al. (2014), Zhang et al. (2015); Chen et al. (2016)). However, in circumstances where one needs knowledge on the specific interconnections of its components, these methods lose their utility. Instead, one requires methods that derive the structure of an entire network from limited, partial, information : i.e. *network reconstruction*.

For the reconstruction of networks based on limited available information, there are several methods which may be used. As described in (Squartini et al. (2018)), many are based on the maximization of Shannon entropy subject to the available information, others are based on similar maximization however using different functionals of entropy, and some offer alternatives to entropy-maximization altogether. The first category, where constrained Shannon entropy is maximized to derive a solution for the desired reconstruction procedure, may be further subcategorized depending on the densities of the resulting reconstructed (or estimated) networks. The first of three subcategories produces networks which are dense, i.e. networks that are (generally too) tightly interconnected, the second produces networks with adjustable densities, and the third produces networks where the exact density of the target network is reproduced in its estimates.

With many existing methods for the reconstruction of networks on the basis of partial, limited information, recent research has paid special attention to comparing their performance. Although the general idea of the papers relating to this topic is to compare methods, the examined methods and data differ.

In Ramadiah et al. (2017), four approaches are used to reconstruct the weighted, directed network for the bank-firm credit market in Japan. Each approach is applied and evaluated in various categories: the ability to recover network characteristics of the original network, the ability to recover the topological structure (i.e. the links) of the original network, and the ability to recover the weight distribution of the original network. Apart from these categories, the systemic risk is also evaluated on the estimated networks. Furthermore, each analysis is observed on three separate levels of aggregation. From this it is clear that there is not a single method that performs best in all the varying categories.

In Anand et al. (2017), seven methods proposed in various papers, both deterministic and probabilistic, are compared. The various approaches are applied to seven types of financial networks spanning twelve different countries and the Eurozone. For method-comparison, two groups of similarity measures are used to evaluate how similar the reconstructed networks are to the original network. The first group focuses on how accurately the reproduced networks capture the links which are either present or absent in the original network, while the second group focuses on how well the reproduced networks capture the size of these links. One further measure is also used to evaluate the systemic risk. Based on the results, the authors devise a few rules of thumb in regards to which methods should be used for which tasks. Similar studies may be found in Mazzarisi and Lillo (2017) and Lebacher et al. (2019).

The general design of these method-comparison studies is that a number of network reconstruction techniques are selected as well as the categories in which they are to “compete” against each other. These categories typically include the recovery of the topological structure of the target network, the recovery of the weight distribution of the target network or the recovery of certain network properties.

In a similar fashion to the aforementioned studies, the present paper will also conduct a comparison of network reconstruction methods. As previously stated, one of the investigated approaches is one developed by Statistics Netherlands for the reconstruction of the Dutch inter-firm transaction transaction on the commodity group level. SN’s reconstruction procedure is based on a few empirically-backed economic observations as well as assumptions made about the link density per commodity group. One goal is to compare the performance of SN’s method with a method well-established in the literature. This comparison will also uncover the biases, if

any, arising from the incorporation of assumptions in the place of using exclusively what is known for certain, i.e. the data at hand. The contribution of the present paper is to examine any differences arising in the methods that do not maximize what is not known for certain, such as the method by SN, and method which do, such as entropy-maximizing methods, when they are fed the exact same input.

1.2 Objectives and Outline

The objectives of the current paper are essentially threefold. The primary goal is to investigate potential biases arising in network reconstruction methods which guide the reconstruction procedure based on realistic assumptions rather than remaining exclusively noncommittal to unknown information. The second goal is to present an adequate design for such a method comparison. A product of the second goal is applying entropy-maximization to the case of the reconstruction of a multiplex. The third and last goal is to evaluate the performance of the entropy-maximizing method with respect to networks of varying densities.

The paper is structured as follows. The second section is a brief overview of some of the basic concepts which are used in the subsequent sections. The third section is dedicated to precisely outlining the methods, which are to be applied and compared to each other, as well as the describing the input data. The fourth section presents and analyzes the outcome of each applied method. In the fifth and final section, relevant conclusions are drawn from the results and discussed. Should the reader desire a more detailed version of the present paper, they may refer to Rachkov (2020).

2 Theoretical Background

The current section gives a brief description of the network formalism and the necessary theoretical background of the maximum entropy framework.

2.1 Networks

A network, $\mathcal{G}$, is a set of points and the relationships between these points or, in more formal terms, a set of nodes and edges. In graph theory, two nodes i and j that are connected to each other through a link e_{ij} are said to be adjacent to each other and incident with the link e_{ij} (Wilson (1996); van Steen (2010)).

The adjacency matrix of a network, $\mathcal{A}(G) = A$, is essentially its binary representation. In the adjacency matrix of a simple, directed network, a 1 in the entry of the i^{th} row and j^{th} column, a_{ij}, specifies a relationship from a node i to a node j, while a 0 indicates no relationship between these nodes. Taking the row sums or the column sums of A produces the out-degrees (i.e. numbers of out-going links) and in-degrees (i.e. numbers of in-going links) of the individual nodes, respectively (Barabási and Pósfai (2016)).

Besides direction, a relationship between nodes may be further defined by a weight (Lerner et al. (2009); Barabási and Pósfai (2016); van Steen (2010); Cimini et al. (2015)). The weight matrix

$\mathcal{W}(G) = W$ may therefore be viewed as a more informative extension of the A, since $w_{ij} > 0$ implies $a_{ij} = 1$ Cimini et al. (2015). Here, the marginals of $\mathcal{W}(G) = W$, i.e. the row and column sums, are called the out- and in-strengths of nodes, respectively. In other words, the out-strength s_i^{out} and in-strength s_i^{in} of a node i is the sum of the link weights going out and coming into that node, respectively.

As a last remark about network formalism, networks may also be multi-layered such that a relationship in one layer of the network indicates a specific subtype of connection between nodes.

2.2 Entropy

Before delving into entropy maximization, it is important to understand the key concept on which the framework is founded, i.e. entropy itself. The more ways in which system components can be arranged such that the macroscopic state is satisfied, the higher the entropy of the system will be. Conversely, if there is merely one possible configuration of the system that would satisfy this state, then the entropy of the system would in fact be zero (De Martino and De Martino (2018)). Entropy, therefore, can be seen as a measure of uncertainty (Robinson (2008); Jaynes (1957)). With this in mind, when one refers to the entropy of a probabilistic system or process describing the possible realizations of a random variable, this implies referring to the uncertainty surrounding this system or process and the "information needed to fully describe it" (i.e. its average information) (Morais (2018)). Intuitively, the more information needed to describe an instance of a variable (i.e. an event), the more uncertain it is and vice versa. Formally, if X is a random variable, then its information content of a realization of X, or alternatively, its degree of surprise is given by $I(x) = -\log[p(x)]$ (Squartini et al. (2018); Morais (2018); Baldi (2002)). One can see that the intuition outlined in the previous sentences does in fact hold with this definition. Indeed, if an event is completely certain, i.e. $p(x) = 1$, then the degree of surprise associated to observing this event happening is zero: $I(x) = -\log(1) = 0$. On the other hand, if an event is practically impossible, i.e. $p(x)$ is virtually 0, then degree of surprise associated to observing this event happening is virtually infinite: $I(x) = -\log(0) = +\infty$ (Squartini et al. (2018)).

To precisely define entropy, this paper will employ the most widely used definition of entropy, which was presented by Claude Shannon and is known as Shannon's entropy (from this point forth, whenever entropy is mentioned, it is in reference to Shannon's entropy). Recall that the entropy of a random variable is the average information over all its possible instances (Morais (2018); Baldi (2002)). With this in mind, Shannon's entropy is therefore given by

$$H(X) = -\sum_{x \in X} p(x) \cdot \log[p(x)] \tag{1}$$

(Squartini et al. (2018); De Martino and De Martino (2018); Morais (2018); Jaynes (1957); Baldi (2002); Baldi (2002)).

There are some basic properties satisfied by entropy. First, entropy is always non-negative $H(X) \geq 0$. Second, it increases with increasing uncertainty, i.e. when $p(x)$ approaches a uniform distribution where no state is more probable than the other states. Third, it "is additive for independent sources of uncertainty" (Jaynes (1957)).

2.3 Entropy Maximization

The maximum entropy framework provides an extremely powerful tool to construct probability distributions of random variables from partial information (Jaynes (1957)). Before tailoring maximum entropy to the task of network reconstruction, the general framework will first be described.

Suppose one were to enforce constraints on the entropy of a system, such that these constraints represented the available information about the system. Then the constrained entropy would capture the multitude of ways in which the system components can be configured such that they all lead to the same constrained state of the system and where the constraints represent what is known for certain.

The idea of the constrained maximization of entropy is proposed and outlined in Edwin Thomas Jaynes' well-cited 1957 paper titled 'Information Theory and Statistical Mechanics' (Jaynes (1957)). In this paper, he describes the problem of assigning probabilities to the outcomes of a variable based on incomplete information. This paper adresses how to devise the probability distribution of a variable, from which one could then make inferences, when faced with limited information, while simultaneously not introducing any biases in the process. The solution is to use that probability distribution which has maximum entropy subject to what is known. Using exactly this probability distribution would imply that one remains "maximally non-committal" to unknown information.

Assuming one has knowledge about M observed properties of the system of interest, one would want to enforce that the sought-for probability distribution will replicate the values of all these M properties as expected values. Here, it is important to make the distinction between microcanonical ensembles and canonical ensembles. In the former, the imposed constraints are "hard" constraints, while in the latter, they are "soft" constraints. Using a hard-constraints approach implies that the resulting (microcanonical) ensemble will satisfy exactly the imposed constraints. On the other hand, using a soft-constraints approach, which is used in this paper, the resulting (canonical) ensemble will satisfy these constraints on average (Squartini et al. (2015)).

Mathematically, the derivation of this desired distribution may be expressed as so. For a random variable X, with the available moment information $\{\langle f_m \rangle = \sum_{x \in X} p(x) \cdot f_m(x)\}_{m=0}^{M}$ for $m = 0, \ldots, M$, one may maximize the entropy of X constrained by this information

$$\begin{aligned} &\underset{p(x)}{\text{maximize}} \quad -\sum_{x \in X} p(x) \cdot \log p(x) \\ &\text{subject to} \quad \langle f_m \rangle = \sum_{x \in X} p(x) \cdot f_m(x) \ \text{ for } \ m = 0, \ldots, M \end{aligned} \tag{2}$$

(Morais (2018)). To solve the above optimization problem, the Lagrangian multipliers method is used (hence, the λ's in the solution). Such an optimization procedure goes about in such a fashion. First, one must construct a Lagrangian

$$\mathcal{L} = -\sum_{x \in X} p(x) \cdot \log p(x) - \sum_{m=0}^{M} \lambda_m \left(\sum_{x \in X} p(x) \cdot f_m(x) - \langle f_m \rangle \right) \tag{3}$$

Squartini et al. (2018); Morais (2018). One must acknowledge that the first constraint is always $\langle f_0 \rangle = 1$ and $f_0(x) = 1$ so that $\sum_{x \in X} p(x) = 1$. Such a constraint is necessary to ensure that the normalization condition is met in order to obtain a valid probability distribution (Squartini et al. (2018); Morais (2018); Jaynes (1957)). Thus, the Lagrangian may be written as

$$\mathcal{L} = -\sum_{x \in X} p(x) \cdot \log p(x) - \lambda_0 \Big(\sum_{x \in X} p(x) - 1 \Big) - \sum_{m=1}^{M} \lambda_m \Big(\sum_{x \in X} p(x) \cdot f_m(x) - \langle f_m \rangle \Big) \qquad (4)$$

To obtain the solution, one must solve the partial derivatives of $\mathcal{L}$ with respect to $p(x)$ and all $\lambda = \{\lambda_m\}_{m=0}^{M}$ (Morais (2018)). The resulting solution is

$$\begin{aligned} p(x) &= \exp\Big(-1 - \lambda_0 - \sum_{m=1}^{M} \lambda_m \cdot f_m(x) \Big) \\ &= \exp\big(-1 - \lambda_0 \big) \cdot \exp\Big(- \sum_{m=1}^{M} \lambda_m \cdot f_m(x) \Big). \end{aligned} \qquad (5)$$

(Squartini et al. (2018); Morais (2018)).

The inverse of the first part of the solution, $\exp\big(1 + \lambda_0\big)$, is in fact the so-called partition function $Z = \sum_{x \in X} \exp\big(- \sum_{m=1}^{M} \lambda_m \cdot f_m(x) \big)$ that acts as a normalizing constant for the probability distribution. Additionally, the other exponent, $\sum_{m=1}^{M} \lambda_m \cdot f_m(x)$, is the so-called Hamiltonian, $H(x)$. The solution for the probability distribution function of X may therefore be rewritten as

$$p(x) = \frac{e^{-H(x)}}{Z} \qquad (6)$$

(Squartini et al. (2018)).

The above expression is very popular in statistical physics, where it is called the Boltzmann-Gibbs distribution. The Hamiltonian is in that case the (measurable) total energy of the system, and the probability distribution is defined over the space of possible (unobservable) microscopic states of the system. Indeed, since the expected total energy is a conserved quantity for a closed system at fixed temperature, in presence of no other piece of empirical information the least biased description of the microscopic states can be obtained by maximizing the entropy subject to a given value of the expected energy.

2.4 Entropy Maximization and Network Reconstruction

The maximum entropy framework outlined above can be applied to many problems. The concrete application this paper will focus on is network reconstruction, specifically, reconstructing the topological structure of a network. The intended goal is thus to reconstruct the adjacency matrix of a network based on partial information. This reconstruction process can be viewed from two perspectives. From the first perspective, it may be seen as an attempt to correctly place the links within a blank adjacency matrix. From the second perspective, it may be

seen as having a collection of unknown numbers and wishing to assign them values, where the unknown numbers are essentially the entries of the adjacency matrix.

Firstly, the random variable to which entropy-maximization will be applied is a network variable $\mathcal{G}$ representing the system of interest. Recall that the system components of a network are the nodes and the links, therefore, making the way in which the links are placed between the nodes a system configuration. As such, each possible realization or outcome of this variable, $G \in \mathcal{G}$, is a possible system configuration given the constraints. In other words, this "ensemble" of networks offers all feasible arrangements of the network that are in agreement with the known information (Squartini et al. (2018)).

Recall the distinction between the microcanonical ensemble of networks and canonical ensemble of networks. In the former, the imposed constraints are "hard" constraints, where the resulting (microcanonical) ensemble will satisfy exactly the imposed constraints. While in the latter, the imposed constraints are "soft" constraints, where the resulting (canonical) ensemble of networks will satisfy these constraints on average (Squartini et al. (2015)). The maximum-entropy methods in the current paper are employed using "soft" constraints and, accordingly, produce canonical ensemble of networks.

Next, if one replaces x with G and X with $\mathcal{G}$ in the solution (6) of the constrained optimization problem, then it produces the expression for the so-called Exponential Random Graphs (ERG) probability distribution:

$$p(G) = \frac{e^{-\sum_{m=1}^{M} \lambda_m \cdot f_m(G)}}{\sum_{G \in \mathcal{G}} \exp\left(-\sum_{m=1}^{M} \lambda_m \cdot f_m(G)\right)} = \frac{e^{-H(G)}}{Z} \tag{7}$$

(Squartini et al. (2018)). Indeed, ERGs can be seen as maximum-entropy ensembles of graphs with given expected properties. ERGs are the root of many proposed reconstruction methods including some of the reconstruction methods utilized in this paper.

With everything else defined, some attention will be paid to the type of available information that may be used to constrain the entropy of $\mathcal{G}$. One option is to constrain the entropy by the in- and out-degree sequences, i.e. $\{k_i^{in}\}_{i=1}^{N}$ and $\{k_i^{out}\}_{i=1}^{N}$, respectively (Parisi et al. (2018)). Note that the expected value of a link existing from node i to node j, $\langle a_{ij} \rangle$, is equal to the probability of a link existing from node i to node j, p_{ij}. Indeed, the solution for this probability, given by

$$p_{ij} = \frac{e^{-\lambda_i^{out}} e^{-\lambda_j^{in}}}{1 + e^{-\lambda_i^{out}} e^{-\lambda_j^{in}}} = \frac{x_i^{out} x_j^{in}}{1 + x_i^{out} x_j^{in}} \tag{8}$$

(Squartini et al. (2018); Parisi et al. (2018)) guarantees that $\langle k_i^{in} \rangle = \sum_{j \neq i} p_{ji}$ and $\langle k_i^{out} \rangle = \sum_{j \neq i} p_{ij}$ since $k_i^{in} = \sum_{j \neq i} a_{ji}$ and $k_i^{out} = \sum_{j \neq i} a_{ij}$. The values x_i^{in} and x_i^{out} come from $P(G) = \frac{e^{-H(G)}}{Z(\lambda^{out}, \lambda^{in})}$ where $H(G) = \sum_i \lambda_i^{out} k_i^{out} + \lambda_i^{in} k_i^{in}$ (Squartini et al. (2018)).

However, having the in- and out-degree sequences of a network is already more information than is realistically available. Without this information, the Lagrange multipliers can not be computed, leaving the solution out of reach. Fortunately, one may make use of a "fitness ansatz",

which asserts that the probability of link existence from one node to another is influenced by intrinsic features of the nodes themselves called fitnesses (Garlaschelli and Loffredo (2004)). This statement can also be referred to as the good-gets-richer phenomenon, similar to the rich-gets-richer phenomenon (Caldarelli et al. (2002)). This mechanism describes a process by which nodes prefer to link themselves to ”higher fitness” nodes. In practice, where it concerns financial and economic networks, it turns out that strengths may successfully be used as fitnesses (Squartini et al. (2018); Cimini et al. (2015)). Furthermore, through empirical evidence it has been shown that strengths (or fitnesses) are correlated with the Lagrange multipliers brought about by constraining entropy by the degree sequences (Parisi et al. (2018); Cimini et al. (2015)). Given this observation, one may transform x_i^{out} and x_i^{in} to $f(x_i^{out}) = \sqrt{a}s_i^{out}$ and $g(x_j^{in}) = \sqrt{b}s_j^{in}$ (Squartini et al. (2018); Parisi et al. (2018)), respectively, in which case:

$$p_{ij} = \frac{\sqrt{a}s_i^{out}\sqrt{b}s_j^{in}}{1 + \sqrt{a}s_i^{out}\sqrt{b}s_j^{in}} \tag{9}$$

$$p_{ij} = \frac{zs_i^{out}s_j^{in}}{1 + zs_i^{out}s_j^{in}} \tag{10}$$

(Squartini et al. (2017); Squartini et al. (2018); Parisi et al. (2018); Cimini et al. (2015)). There is now an unknown parameter $z = \sqrt{ab}$, which has yet to be determined. If one has available the number of links in the network L, then z may be computed in such a way that the expected number of links $\langle L \rangle$ is equal to this observed value, i.e. $L = \langle L \rangle$. Since $L = \sum_i \sum_{j \neq i} a_{ij}$ and therefore $\langle L \rangle = \sum_i \sum_{j \neq i} \langle a_{ij} \rangle = \sum_i \sum_{j \neq i} p_{ij}$, z can be fixed by solving the following equation:

$$L = \sum_i \sum_{j \neq i} \frac{zs_i^{out}s_j^{in}}{1 + zs_i^{out}s_j^{in}} \tag{11}$$

(Squartini et al. (2018); Parisi et al. (2018)). Once z is computed one may use the probability $p_{ij} = \frac{zs_i^{out}s_j^{in}}{1+zs_i^{out}s_j^{in}}$ to establish whether two nodes i and j have a link in the reconstruction process using the following rule:

$$\hat{a}_{ij} = \begin{cases} 1 & \text{with probability } p_{ij} \\ 0 & \text{with probability } (1 - p_{ij}) \end{cases} \tag{12}$$

Hence, with limited information, namely, the adjacency matrix marginals $\{s_i^{out}\}_{i=1}^N$ and $\{s_i^{in}\}_{i=1}^N$ as well as the number of links L, one may reconstruct the desired network. That being said, although the strengths are usually easy to obtain, sometimes the number of links is still unavailable. However, if one had access to a sample of the network, which included a subset of nodes $\mathcal{S}$ of size $n_s = |\mathcal{S}|$, then they could determine z based on the density of this portion of the network, c_s. Therefore, one would solve the following equation relating the observed density of the sample to the expected density:

$$c_s = \frac{1}{n_s(1-n_s)} \sum_{i \in \mathcal{S}} \sum_{j \neq i \in \mathcal{S}} \frac{z s_i^{out} s_j^{in}}{1 + z s_i^{out} s_j^{in}} \tag{13}$$

(Squartini et al. (2017)). Following this, the network reconstruction process is exactly the same as above.

3 Overview of the methods being compared

With the earlier outlined objectives in mind, there are three network reconstruction techniques that are set to the task of reconstructing an inter-firm transaction multiplex, where the layers represent different commodity groups. In order to operationalize the intended research goal of uncovering potential biases in network reconstruction methods such as the one developed by Statistics Netherlands (see Section 3.2), two methods are derived within the maximum entropy framework for the purpose of comparison (see Section 3.3). The precise reason for which these methods are derived within this particular framework is because, as previously explained in Section 2.2.2, entropy-maximization techniques remain maximally non-committal to unknown information, thus, producing the least possible biased network estimates. The first such method is to be calibrated on the number of estimated links by SN's method for each commodity group. Hence, it maximizes uncertainty with respect to the known information with the exception of the implied number of links. The second method is purely least biased as it calibrates its reconstruction procedure on the links on a known portion of the commodity group network.

Furthermore, all methods are to have the exact same input data (see Section 3.1). These data consists of the marginal information on firm-level sales and purchasing volumes (in euros) derived by SN from real industry-level data. In addition to the sales and purchasing volumes, firm coordinates also constitute the input data so as to incorporate distances between firms in the reconstruction process. The last piece of data fed to all methods is the input-output information based on the economic classification of firms.

As the broad idea is to compare the aforementioned methods to each other in terms of the types of networks they reconstruct as well as their performance, diverse comparison measures are used (see Section 3.4). To examine the resulting network estimates, various network properties are measured. To test the performances of each method, in terms of their ability to recover the known portions of the examined commodity group networks, various performance indicators are measured, such as the ability of a method to correctly place present links or absent links.

3.1 Network and Data Description

As previously stated, the network to be reconstructed from partial information is essentially a multiplex, specifically, the Dutch inter-firm transaction network broken down to the commodity group level. Due to computational limitations, a simplified version of this multiplex is used where

only 4 commodity group levels are considered, namely, the commodity groups for government services concerning the environment, water, telecommunication services, and aviation services. There is no particular reason to select these specific commodity groups other than that they constitute the least amount of possible links across all commodity groups and thus require the least amount of computation time. Taking the notation from Section 2.1.4 this implies that $\alpha = 1, \ldots, 4$ and that the adjacency matrix of the inter-firm transaction network is composed of the adjacency matrices of all the considered commodity group networks, $A = (A^{[1]}, \ldots, A^{[4]})$. Here, the rows represent the supplying firms (i.e. the firms supplying some commodity) and the columns represent the using firms (i.e. a firm using a given commodity).

Furthermore, if each $A^{[\alpha]}$ were enriched with weights such that it became $W^{[\alpha]}$, then each entry $w_{ij}^{[\alpha]}$ would be the sales or purchasing volumes between two firms i and j. In other words, it would represent the amount a supplier i sold of a particular commodity to user j (quantified in euros) or, equivalently, the amount a user j purchased of a particular commodity from supplier i. Although this particular information is unavailable, the marginal information (i.e. the in- and out-strengths $s_i^{out,[\alpha]}$ and $s_j^{in,[\alpha]}$) is. In Hooijmaaijers and Buiten (2019), the pair of SN authors present how these firm-level marginals are derived from their corresponding industry-level marginals. Briefly, if the volume of a product (in euros) sold (or purchased) by an industry in a specific commodity group is $D^{out,[\alpha]}$ (or $D^{in,[\alpha]}$), then the volume of the product sold (or purchased) by a firm i in that specific commodity group, i.e. its out-strength (or in-strength) in the commodity group network, is calculated by taking a specific proportion of this amount. The proportion is defined by the net turnover of the firm i over the total net turnover in its industry. This amounts to

$$\begin{aligned} v_i^{out,[\alpha]} &= \frac{\text{net turnover firm } i}{\text{total net turnover industry}} \cdot D^{out,[\alpha]} \\ v_i^{in,[\alpha]} &= \frac{\text{net turnover firm } i}{\text{total net turnover industry}} \cdot D^{in,[\alpha]} \end{aligned} \tag{14}$$

Note that the out- and in-strengths representing sales/purchasing volumes of a firm i in commodity group network α will be denoted $v_i^{out,[\alpha]}$ and $v_i^{in,[\alpha]}$ from this point onwards. To make these values even more realistic some further processing is applied. For a detailed explanation of the derivation procedure refer to Hooijmaaijers and Buiten (2019).

Besides the strengths, additional information incorporated into the reconstruction procedures include distances and input-output relationships. The distance data is obtained from some (x,y)-coordinates be they the latitudinal and longitudinal coordinates of firms or the Rijksdriehoek (RD) coordinates of firms. Furthermore, the input-output information is based on the firm NACE codes (i.e. the codes used in the European industry standard classification system). This means each firm has a given code depending on how they are classified within this system. The data on the existing binary input-output relationships based on these codes is also available for the purpose of reconstruction.

A last and vital piece of available information is the data on existing links in varying portions of the commodity group networks. This information is essential for two reasons: (1) for the calibration of one of the entropy-maximizing methods; (2) to test the performance of all methods in regards to their ability to recover those known portions of the networks. Furthermore, these known network portions represent all the out-going links for a subset of the supplying firms in a

given commodity group. For illustration, this is the same as having complete information on a full horizontal band of the network adjacency matrix.

3.2 Statistics Netherlands Method

The method developed by Statistics Netherlands for the estimation of the Dutch inter-firm transaction network Hooijmaaijers and Buiten (2019) is a deterministic one as the resulting networks per commodity group are constructed without the incorporation of any probabilistic elements. Therefore, the reconstruction procedure outputs only one network per commodity group and, in repeats of the procedure, the resulting networks will always be the same.

Furthermore, the manner in which links are assigned between a using firm and a supplying firm are based on scores derived from several firm features, namely, the sales volumes, geographical distances and input-output relationships (at the level of NACE industry groups). These features as well as the computation of scores are explained with greater detail in Section 3.2.1.

Once the scores are computed, it is then possible to reconstruct the network per commodity group as follows. For each using firm, in the order from largest to smallest purchasing volume, the top t most likely suppliers have their out-going links assigned to the using firm, where the value of t is equal to the using firm's degree (i.e. the in-degree). The firms which are considered to be the most likely suppliers (i.e. the ones with a high likelihood of trading with the using firm in question) are those with the highest scores. Additionally, each of these supplying firms must of course also have remaining out-going links, otherwise, the next most likely supplier is considered. This reconstruction procedure is more thoroughly outlined in Section 3.2.3.

3.2.1 Scores and Parameter Settings

As already mentioned, the reconstruction procedure proposed by Statistics Netherlands involves the assignment of out-going links from supplying firms to the using firms on the basis of score values. These scores represent the combination of three separate components: a company score, a distance score and an industry score. Before any reconstruction procedure may take place, it is necessary to first compute these scores from the data.

Company Score

The first component of the overall score is the company score, which is derived from the out-going volume (i.e. the sales volume) and computed separately for each firm per commodity group. Each company score represents a value for one firm relative to all the other firms in the same commodity group. It thus takes on a value between 0 and 1 and is calculated as follows

$$\delta_i^{[\alpha]} = \max_i[\log(v_i^{out,[\alpha]})] - \log(v_i^{out,[\alpha]}) \tag{15}$$

$$\text{company score}_i^{[\alpha]} = 1 - \frac{\delta_i^{[\alpha]}}{\max_i(\delta_i^{[\alpha]})} \tag{16}$$

where v denotes the sales volume, i denotes the supplying firm, and $\max_i$ denotes maximum values across all supplying firms in a commodity group. From the definition above, it is clear that there will always be a firm with company score$_i^{[\alpha]} = 1$. This is the firm with the maximum sales volume in that commodity group.

It should further be mentioned that defining the company score in such a way implies that larger firms prefer to link to each other. This is due to the fact that during the reconstruction procedure, larger using firms are the first to be linked with supplying firms and will thus be linked to the firms with the highest company scores, which are essentially also the largest supplying firms.

Distance Score

The next component constituting the overall score is the distance score, which is derived from the latitudinal and longitudinal coordinates of firms. For each using firm j, the distance between it and each possible supplier i is calculated as follows

$$d_{ij} = \text{abs}(\text{x}_i - \text{x}_j) + \text{abs}(\text{y}_i - \text{y}_j) \tag{17}$$

$$\text{distance score}_{ij} = \frac{d_{ij}}{\max_i d_{ij}} \tag{18}$$

where x denotes the x-coordinate (in this case, the latitudinal coordinate) and y denotes the y-coordinate (in this case, the longitudinal coordinate). Note that for computational ease in the actual reconstruction procedure, $\max_i d_{ij}$ is replaced with the sum of the maximum absolute differences of the latitudinal and longitudinal-coordinates.

Once more, the definition of the distance score is such that it takes on a value between 0 and 1 and there will always be a supplier, this time per user, which has a distance score$_{ij} = 1$. This is the supplying firm that has the maximum distance from the using firm.

Moreover, as it is later subtracted in the weighted score, the distance score implies that using firms are less likely to trade with supplying firms that are geographically located further away. SN incorporates such an assumption based on empirical evidence supporting the important role geographical distance plays in the choice of supplier, even when a country is small.

Industry Score

The last component for the overall score is the industry score, which is derived from the NACE input-output relationship data. The value serves as a bonus or penalty contingent on whether the using firm's industry actually trades with the supplying firm's industry or not.

$$\text{industry score}_{ij} = \begin{cases} 0.1 & \text{when NACE groups of supplier } i \text{ and user } j \text{ do trade} \\ -1 & \text{when NACE groups of supplier } i \text{ and user } j \text{ do not trade} \end{cases} \tag{19}$$

Combining Scores

The actual reconstruction procedure, where networks are produced by linking using and supplying firms to each other for each of the commodity groups, utilizes a combination of the above-described scores. The company score and the distance score are assigned weights such that the weights add up to 1. Therefore, these weights complement each other whereby the weight of the distance score is one minus the weight of the company score, β. As last, the industry score is added to the weighted sum of the company and distance scores. The overall score for each supplier i with respect to a given supplier j in some commodity group is obtained by the following

$$\text{score}_{ij}^{[\alpha]} = \beta \cdot \text{company score}_{i}^{[\alpha]} + (1-\beta) \cdot (1 - \text{distance score}_{ij}) + \text{industry score}_{ij} \quad (20)$$

The weight β essentially determines the relative importance of company size and geographical distance. Although a different value for this weight can be chosen for each commodity group (since not every type of good or service depends on geographical distance in the same way), the same value is still used across all commodity groups. Specifically, this value is chosen to be $\beta = 0.7$.

3.2.2 Estimation of In- and Out-Degrees

Besides score values, the reconstruction procedure proposed by SN additionally requires the number of in-going links for each using firm and the number of out-going links for each supplying firm in each commodity group. Their approach to estimating the in-degrees and out-degrees is briefly explained here.

In-Degrees

First, the in-degrees of each using firm is calculated in each commodity group α. This is done in the following fashion for each user j

$$k_j^{in,[\alpha]} = \left[\log(v_j^{in,[\alpha]}) - \min_j [\log(v_j^{in,[\alpha]})] + 1 \right]^{\eta} \quad (21)$$

where η is an adjustment factor, which is assumption-based and chosen to be $\eta = 0.5$. Furthermore, $k_j^{in,[\alpha]}$ is rounded to the nearest full number since degrees are discrete.

Out-Degrees

Next, the out-degrees per firm in each commodity group are computed using the fact that $k^{in,[\alpha]} = \sum_j k_j^{in,[\alpha]} = k^{out,[\alpha]}$ as well as an observation taken from Watamabe et al. (2013), namely, that the turnover of a firm i and its degree share a power-law relationship. This relationship is defined by

$$\text{turnover}_i = \Gamma \cdot k_i^{\gamma} \quad \Leftrightarrow \quad k_i = \left(\frac{\text{turnover}_i}{\Gamma} \right)^{\frac{1}{\gamma}} \quad (22)$$

where $\gamma = 1.3$ (note that this power-law relationship is in contrast with the aforementioned "fitness ansatz" in the maximum entropy method, where a different nonlinear relation between turnover and degree is used; indeed, this is an important difference between the two methods). Furthermore, an initial estimate of the constant Γ may be computed as $\Gamma_0 = \left(\frac{\sum_i \text{turnover}_i^{1/\gamma}}{\sum_i k_i} \right)^{\gamma}$. To account for the fact that Γ and the degrees do not have a continuous relationship and thus to ensure a valid estimate for Γ, two techniques are implemented: bracketing and bisection. For more details refer to Hooijmaaijers and Buiten (2019).

To sum up, the out-degrees of the supplying firms in each commodity group α can therefore be estimated using the initial estimate $\Gamma_0^{[\alpha]} = \left(\frac{\sum_i \text{turnover}_i^{1/1.3}}{k^{out,[\alpha]}} \right)^{1.3}$ to find $\Gamma^{[\alpha]}$ and subsequently using the relationship $k_i^{out,[\alpha]} = \left(\frac{\text{turnover}_i}{\Gamma^{[\alpha]}} \right)^{1/1.3}$.

3.2.3 Reconstruction Procedure

With all the necessary components explained, it is now possible to detail SN's reconstruction procedure (for pseudocode refer to Algorithm 1). For each commodity group $\alpha = 1, \ldots, C$, in which there are $N_u^{[\alpha]}$ unique users and $N_s^{[\alpha]}$ unique suppliers, links are assigned based on the following steps.

First, users are ordered from largest to smallest purchasing volume. The first user will therefore be the one with the largest purchasing volume in the commodity group. Furthermore, this user has an in-degree $k_1^{in} = t$, which is essentially the number of suppliers from which it will have in-coming links. As explained earlier, these suppliers are selected on the basis of the computed scores (recall that these scores are computed per user for all suppliers). Thus, picking the top t suppliers the user is most likely to trade with means picking the t suppliers having the highest scores. Each of the selected suppliers i then have 1 subtracted from their out-degrees, $k_i^{out} \leftarrow k_i^{out} - 1$.

Once the first links are assigned, the procedure carries on to the next users. The same logic applies, however, as the algorithm progresses it might be that one or more of the top t suppliers for the current user do not have any leftover out-going links. In this case, the suppliers with the next highest scores are chosen. The process continues until all links have been assigned.

Algorithm 1: Statistics Netherlands Network Reconstruction Procedure

Data: input data per commodity group

Result: directed network per commodity group

begin

  for $\alpha = 1, \ldots, C$ **do**

    order users from largest to smallest transaction volume

    for $j = 1, \ldots, N_u^{[\alpha]}$ **do**

      $t \leftarrow k_j^{in}$

      select subset T of top t suppliers based on highest scores having nonzero out-degrees

      for $i \in T$ **do**

        assign link $i \rightarrow j$

        $k_i^{out} \leftarrow k_i^{out} - 1$

      end

    end

  end

end

3.3 Maximum Entropy Methods

To compare the performance of SN's method and to investigate what type of biases may arise in the resulting networks due to the incorporation of various assumptions, two nearly identical entropy-maximizing methods are derived such that they make use of precisely the same data as SN's method. The difference between them is the number of links on which their procedures are calibrated. The first entropy-maximizing method, named ME Method 1, is calibrated on the number of links estimated by SN. In essence, this means that it will not produce the least biased results as it is maximally non-committal to what is unknown with the exception of the incorporation of one assumption made by SN. On the other hand, the second entropy-maximizing method, named ME Method 2, is calibrated on the number of links in a known portion of the network. As such, the second method is purely entropy-maximizing with respect to what is known for certain.

3.3.1 Matching the Input Data

To ensure a valid and sound comparison between the methods, the pieces of input data should be the same, namely, the sales and purchasing volumes of firms (i.e. the out- and in-strengths of the network nodes), distances between firms, and the binary input-output relationships of firms.

Sales and Purchasing Volume

As explained in Section 2.2.3, the in- and out-degree sequences are most times not the type of information one has available. Thanks to the fitness ansatz however, if the number of links as well as the in- and out-strengths are known, then it is possible to obtain a solution for the probability of link existence between two nodes i and j. The strengths used are therefore the in- and out-going volumes (or the Use and Supply, respectively) of the $N_u^{[\alpha]}$ using firms and the $N_s^{[\alpha]}$ supplying firms in each commodity group α, $\{v_i^{in,[\alpha]}\}_{i=1}^{N_u^{[\alpha]}}$ and $\{v_i^{out,[\alpha]}\}_{i=1}^{N_s^{[\alpha]}}$, respectively. Hence, by the good-gets-richer mechanism defining the fitness ansatz, the higher the sales/purchasing

volume of a firm, the higher the probability of other firms trading with it. Note that these volumes are not normalized by the maximum volume like in the SN method and are instead used in their raw form in both entropy-maximizing reconstruction procedures.

Distance

The next piece of required data is the distance between firms. Contrary to SN's method, the coordinates used to compute these distances are the Rijksdriehoek (RD) coordinates. Although the type of coordinates are not precisely the same, this is not expected to make a difference for the interpretation of the results.

Furthermore, the coordinates are handled in roughly the same way as with the SN method. Again, the sum of the absolute differences of the x- and y-coordinates are taken and normalized by a maximum distance. In this case, the maximum distance is also taken to be the sum of the maximum absolute differences between the x and y-coordinates. The distance between a supplier i and user j is therefore given by

$$d_{ij} = \frac{\text{abs}(\text{x}_i - \text{x}_j) + \text{abs}(\text{y}_i - \text{y}_j)}{\text{max distance}} \tag{23}$$

As the distances in the SN method, the distances here also take values in the range between 0 and 1.

Input-Output Relationships

The final piece of information to be incorporated into the entropy-maximization methods is the data on the binary input-output relationships. Again, the binary information is based on the NACE codes of firms and this time the value merely indicates whether or not the supplying firm's industry and the using firm's industry actually trade with each other

$$I_{ij} = \begin{cases} 1 & \text{when NACE groups of supplier } i \text{ and user } j \text{ do trade} \\ 0 & \text{when NACE groups of supplier } i \text{ and user } j \text{ do not trade} \end{cases} \tag{24}$$

3.3.2 Link Probabilities

With the same building blocks used to define the reconstruction procedure of SN's method, it is possible to define the reconstruction procedures of the entropy-maximizing methods as well. Recall that due to the empirically observed correlation between fitnesses (which may be represented by strengths) and the Lagrangian multipliers produced by constraining entropy by the degree sequences, the probability of link existence between two nodes may be expressed as $p_{ij} = \frac{z s_i^{out} s_j^{in}}{1 + z s_i^{out} s_j^{in}}$. This requires the estimation of the unknown parameter z, which is the aspect defining the difference between both maximum entropy methods. As one will see in the following subsections, the probabilities, and thus the unknown parameter z, will be commodity dependent.

Maximum Entropy Method 1 (ME Method 1)

The first method constructed within the maximum entropy framework, whose results are to be compared to those produced by SN's method, maximizes uncertainty with respect to what is known with the exception of the implied number of links. Indeed, ME Method 1 calibrates on the number of links estimated by SN, denoted as $L_{SN}^{[\alpha]}$, for the portion of firms in the subset $\mathcal{S}^{[\alpha]}$ of each commodity group network α for which the links are known. This is contrary to ME Method 2 where the considered number of links is the real number of known links for firms in $\mathcal{S}^{[\alpha]}$. As previously mentioned, the known portion of each commodity group network contains all the out-going links for a subset of $n_s^{[\alpha]}$ supplying firms. Thus, $L_{SN}^{[\alpha]}$ is calculated by summing over all the estimated degrees for the supplying firms in $\mathcal{S}^{[\alpha]}$, i.e.

$$L_{SN}^{[\alpha]} = \sum_{i=1}^{n_s^{[\alpha]}} \hat{k}_i^{out,[\alpha]} \tag{25}$$

The estimated densities of these subnetworks are therefore given by

$$c_{SN}^{[\alpha]} = \frac{L_{SN}^{[\alpha]}}{\text{\# possible links in } \mathcal{S}^{[\alpha]}} \tag{26}$$

Here, the denominator represents the total number of possible links excluding self-loops and is computed by taking the product of the number of supplying firms and the number of using firms in $\mathcal{S}^{[\alpha]}$, i.e. $n_s^{[\alpha]} \cdot N_u^{[\alpha]}$, and subtracting the number of firms listed as both a supplier and a user.

Since the number of combinations between the supplying and using firms in the known portions of the networks is very large for every commodity group and poses a computational challenge for estimating the unknown parameter $z^{[\alpha]}$, a random sample of $n_u^{[\alpha]}$ using firms (roughly equal to $0.25 \cdot N_u^{[\alpha]}$) is taken such that the utilized portion of the network is reduced from $\mathcal{S}^{[\alpha]}$ to $\mathcal{S}_{samp}^{[\alpha]}$. Although the number of estimated links in this reduced portion is unknown, it can be assumed that its density is the same as the non-reduced portion. In other words, it can be assumed that its density is equal to $c_{SN}^{[\alpha]}$.

Finally, assembling all the pieces to construct the probability that is used to determine the existence of links in the reconstructed networks produced by ME Method 1, yields the following

$$p_{ij}^{1,[\alpha]} = \frac{(z^{[\alpha]} v_i^{out,[\alpha]} v_j^{in,[\alpha]} I_{ij})/d_{ij}}{1 + (z^{[\alpha]} v_i^{out,[\alpha]} v_j^{in,[\alpha]} I_{ij})/d_{ij}} \tag{27}$$

where the unknown parameter $z^{[\alpha]}$ is found by solving $c_{SN}^{[\alpha]} = \langle c_{SN}^{[\alpha]} \rangle$, i.e.

$$c_{SN}^{[\alpha]} = \frac{1}{\text{\# possible links in } \mathcal{S}_{samp}^{[\alpha]}} \sum_{i=1}^{n_s^{[\alpha]}} \sum_{j=1}^{n_u^{[\alpha]}} \frac{(z^{[\alpha]} v_i^{out,[\alpha]} v_j^{in,[\alpha]} I_{ij})/d_{ij}}{1 + (z^{[\alpha]} v_i^{out,[\alpha]} v_j^{in,[\alpha]} I_{ij})/d_{ij}} \tag{28}$$

It is important to note that the utilized ansatz, $z^{[\alpha]}v_i^{out,[\alpha]}v_j^{in,[\alpha]}I_{ij}/d_{ij}$, is defined as such in order to match the information used in the SN Method and is not explicitly verifiable (in other words, has not been verified in the known portion of the network).

Maximum Entropy Method 2 (ME Method 2)

The second entropy-maximizing method, ME Method 2, entirely maximizes uncertainty with respect to what is known. Hence, it is calibrated on the sum of real degrees in the place of estimated ones. For both methods, the same reduced portion of $\mathcal{S}^{[\alpha]}$ is used, i.e. $\mathcal{S}_{samp}^{[\alpha]}$, which consists of the same $n_s^{[\alpha]}$ suppliers and the same $n_u^{[\alpha]}$ users. Suppose there are $L_{samp}^{[\alpha]}$ known links for the firms in $\mathcal{S}_{samp}^{[\alpha]}$, then the density is given by

$$c_{samp}^{[\alpha]} = \frac{L_{samp}^{[\alpha]}}{\text{\# possible links in } \mathcal{S}_{samp}^{[\alpha]}} \tag{29}$$

Following the same logic as ME Method 1, ME Method 2 defines its probability for link existence as

$$p_{ij}^{2,[\alpha]} = \frac{(z^{[\alpha]}v_i^{out,[\alpha]}v_j^{in,[\alpha]}I_{ij})/d_{ij}}{1 + (z^{[\alpha]}v_i^{out,[\alpha]}v_j^{in,[\alpha]}I_{ij})/d_{ij}} \tag{30}$$

where the unknown parameter $z^{[\alpha]}$ is found by solving $c_{samp}^{[\alpha]} = \langle c_{samp}^{[\alpha]} \rangle$

$$c_{samp}^{[\alpha]} = \frac{1}{\text{\# possible links in } \mathcal{S}_{samp}^{[\alpha]}} \sum_{i=1}^{n_s^{[\alpha]}} \sum_{j=1}^{n_u^{[\alpha]}} \frac{(z^{[\alpha]}v_i^{out,[\alpha]}v_j^{in,[\alpha]}I_{ij})/d_{ij}}{1 + (z^{[\alpha]}v_i^{out,[\alpha]}v_j^{in,[\alpha]}I_{ij})/d_{ij}} \tag{31}$$

3.3.3 Reconstruction Procedure

The two entropy-maximizing methods ME Method 1 and ME Method 2 estimate the unknown parameter in slightly different ways, however, their reconstruction procedures follow precisely the same logic (see Algorithm 2[1)]). As both methods are probabilistic, they output an ensemble of networks since every repeat of the reconstruction procedure does not produce the exact same network. Recall the idea is that the ensemble of networks will reproduce the imposed constraints on average. Thus, the desired size of this ensemble B must also be specified. In this case, the desired size is set to $B = 25$. With 4 commodity groups, this implies that the size of the total produced ensemble across all the layers in the multiplex is 100.

The reconstruction procedures may be described by the following recipe. For each commodity group α, the unknown parameter $z^{[\alpha]}$ is determined depending on the method being used (see

1) As with all code regarding the current paper, the code for both algorithms was written using the statistical language and environment R (R Core Team (2013)). As no code is provided in the Appendix, it can be made available upon request via contact with andrea.rachkov@yahoo.nl

previous sections). Then for each supplier-user pair (excluding self-loops), calculate the probability of a link existing between them, i.e. p_{ij}^1 for Method ME 1 and p_{ij}^2 for Method ME 2. With this given probability, one may then draw from the Bernoulli distribution. If the outcome is a 1, a link is assigned from supplier i to user j. This process is repeated B times.

Algorithm 2: Maximum Entropy Method 1 and Method 2 Network Reconstruction Procedure

Data: input data and known links per commodity group, desired size of ensemble B
Result: ensemble of directed networks per commodity group
begin
  for $\alpha = 1, \ldots, C$ **do**
    take random sample $\mathcal{S}_{samp}^{[\alpha]}$ of known portion of network
    ME Method 1: calculate $c_{SN}^{[\alpha]}$ and solve $c_{SN}^{[\alpha]} = \langle c_{SN}^{[\alpha]} \rangle$ for $z^{[\alpha]}$
    ME Method 2: calculate $c_{samp}^{[\alpha]}$ and solve $c_{samp}^{[\alpha]} = \langle c_{samp}^{[\alpha]} \rangle$ for $z^{[\alpha]}$
    for $b = 1, \ldots, B$ **do**
      for $j = 1, \ldots, N_u^{[\alpha]}$ **do**
        for $i = 1, \ldots, N_s^{[\alpha]}$ **do**
          if *firm* $i \neq$ *firm* j **then**
            ME Method 1: calculate p_{ij} as p_{ij}^1
            ME Method 2: calculate p_{ij} as p_{ij}^2
            assign link $i \rightarrow j$ with probability p_{ij}
          end
        end
      end
    end
  end
end

3.4 Comparison Measures

3.4.1 Estimated Network Properties

To compare the networks produced by each method, several of the network properties outlined in Section 2.1.2 are estimated in the reconstructed networks. These include basic properties such as the total number of links, the maximum degree, the average degree and the minimum degree. In addition, the differences in the shapes of the degree distributions of the reproduced networks (including their power-law distribution parameter estimates) are also contrasted. The out-degree distributions for the known portions of the commodity group networks are also compared to the empirical one using the discrete Kolmogorov-Smirnov statistic, which measures the maximum difference between the distributions[1)]. Furthermore, properties such as the diameter, the global and average clustering coefficients are also examined. Note that for each commodity group layer, these properties are calculated considering only the suppliers and users participating in that commodity group.

[1)] The `ks.test()` function was used from the dgof package in R

3.4.2 Estimated Links

To compare the performances of each method, a few useful indicators are measured. These indicators say something about the ability of each method to recover the links in the known portions of each commodity group network $\mathcal{S}^{[\alpha]}$; in other words, the known subnetworks of each layer in the multiplex. In particular, these indicators are derived from the number of correctly and falsely estimated links.

A link, which exists in the reconstructed subnetwork as well as the observed subnetwork, is called a true positive. The sum of these types of links is denoted here as $TP_{\mathcal{S}^{[\alpha]}}$. Dividing this number by the real number of links in this subnetwork, $L_{\mathcal{S}^{[\alpha]}}$, yields the true positive rate given by

$$TPR_{\mathcal{S}^{[\alpha]}} = \frac{TP_{\mathcal{S}^{[\alpha]}}}{L_{\mathcal{S}^{[\alpha]}}} \tag{32}$$

On the other hand, dividing this number by the estimated number of links in this subnetwork, $\hat{L}_{\mathcal{S}^{[\alpha]}}$, yields the positive predicted value (otherwise known as the precision) given by

$$PPV_{\mathcal{S}^{[\alpha]}} = \frac{TP_{\mathcal{S}^{[\alpha]}}}{\hat{L}_{\mathcal{S}^{[\alpha]}}} \tag{33}$$

Furthermore, a link, which does not exist in either the reconstructed subnetwork or the observed one, is called a true negative. The total of such links is denoted as $TN_{\mathcal{S}^{[\alpha]}}$. Dividing this number by the difference between the total number of possible links and the true number of links in the commodity group subnetwork defines the true negative rate (otherwise known as the specificity) given by

$$SPC_{\mathcal{S}^{[\alpha]}} = \frac{TN_{\mathcal{S}^{[\alpha]}}}{\#\text{ possible links in } \mathcal{S}^{[\alpha]} - L_{\mathcal{S}^{[\alpha]}}} \tag{34}$$

Recall that the number of possible links in a (sub)network where there are N nodes is equal to $N(N-1)$, however, as mentioned in the previous sections, in the case of the commodity group subnetworks, the number of possible links is taken to be the product of the number of supplying firms and the number of using firms in the known portion $\mathcal{S}^{[\alpha]}$, i.e. $n_s^{[\alpha]} \cdot N_u^{[\alpha]}$, and subsequently subtracting the number of firms listed as both a supplier and a user so as to exclude self-loops.

The last indicator, accuracy, takes into account both the links that are correctly estimated as present and the ones correctly estimated as absent. It is therefore computed by taking the sum of the true positives and negatives and dividing it by the total number of possible links in the commodity group subnetwork. This is given by

$$ACC_{\mathcal{S}^{[\alpha]}} = \frac{TP_{\mathcal{S}^{[\alpha]}} + TN_{\mathcal{S}^{[\alpha]}}}{\#\text{ possible links in } \mathcal{S}^{[\alpha]}} \tag{35}$$

Two measures that were not yet mentioned are the false positives and negatives. False positives are the number of links, denoted here as $FP_{\mathcal{S}^{[\alpha]}}$, existing in the reconstructed subnetwork but not

in the observed subnetwork. Opposite to these are the false negatives, denoted as $FN_{\mathcal{S}[\alpha]}$, which are the links not present in the reconstructed subnetwork but are present in the observed one.

Note that the sum of the true positives and false negatives is equal to the number of true links, i.e. $L_{\mathcal{S}[\alpha]} = TP_{\mathcal{S}[\alpha]} + FN_{\mathcal{S}[\alpha]}$, and that the sum of the the true positives and false positives is equal to the number of estimated links $\hat{L}_{\mathcal{S}[\alpha]} = TP_{\mathcal{S}[\alpha]} + FP_{\mathcal{S}[\alpha]}$. Furthermore, the total number of possible links is equal to the sum of all the above mentioned measures, i.e.
possible links in $\mathcal{S}^{[\alpha]} = TP_{\mathcal{S}[\alpha]} + TN_{\mathcal{S}[\alpha]} + FP_{\mathcal{S}[\alpha]} + FN_{\mathcal{S}[\alpha]}$.

3.5 Varying Density Performance

As mentioned in the Introduction, the primary goal of the present paper is to uncover biases, if any, in methods which are not based on entropy-maximization. A secondary goal was also declared, namely, that of investigating the performance of the purely entropy-maximzing method with respect to networks of varying densities. To achieve said goal, a slighty modified version of ME Method 2 is applied to all the known commodity group subnetworks thus resulting in 575 data observations. In this case, the using firms which do not have a known link are excluded from $\mathcal{S}^{[\alpha]}$, implying that each subnetwork incorporates only the n_s supplying firms and the n_u using firms which have known links. The subset of each commodity group is therefore denoted as $\mathcal{S}^{[\alpha]}_{known}$ instead, where the total number of links is denoted as $L^{[\alpha]}_{known}$. Note that the $\alpha = 1, \ldots, 575$. Furthermore, since the supplier-user combinations are much smaller in the subnetworks (which are essentially treated as complete networks), the method is directly calibrated on the links in the place of the density. This means that the unknown parameters $z^{[\alpha]}$ in each commodity group are found by solving $L^{[\alpha]} = \langle L^{[\alpha]} \rangle$.

To explore the effect of varying density on reconstruction performance, the density of each subnetwork is first calculated as

$$c^{[\alpha]}_{known} = \frac{L^{[\alpha]}_{known}}{\text{\# possible links in } \mathcal{S}^{[\alpha]}_{known}} \tag{36}$$

Then, the reconstruction process for ME Method 2 described in Section 3.3.3 is applied to all 575 subnetworks. With the ensemble of reconstructed networks for each commodity group, the measures described in Section 3.4.2 are calculated for each reconstructed network.

4 Results

4.1 Method Comparison Results

From several results, which may be found in Rachkov (2020), there are a few key points worth mentioning in regards to both the behaviors of the studied methods as well as the biases arising in the non-entropy-maximizing method, i.e. the SN Method, with respect to the the entropy-maximizing methods (or the least biased methods), i.e. ME Method 1 and ME Method 2.

For clarity, one should recall that the SN Method is deterministic and, therefore, always produces one solution for each of the estimated networks, while the maximum entropy methods are probabilistic and, therefore, yield various realizations of the estimated networks.

To start, in every commodity group network layer, the number of reconstructed links in the known subportion are the same for the SN Method and ME Method 1, while ME Method 2 reconstructs a number of links equal to the known number of links in the subnetwork with the exception of commodity group 5100000 (the commodity group for aviation services). It should be noted, however, that this exception does not indicate a flaw of ME Method 2, but rather sheds light on the fact that numerical issues may still cause discrepancies between the expected number of links and the actual number of links in the reconstructed networks.

Furthermore, as one would expect to be the case, the more similar the densities on which the maximum entropy methods are calibrated, the more similar their performance in recovering the link structure of the original network are as well the structures of their resulting networks. In cases where the calibration densities are similar for ME Method 1 and ME Method 2, the $TPR_{\mathcal{S}^{[\alpha]}}$ values and the $PPV_{\mathcal{S}^{[\alpha]}}$ values are in sync (or at least relatively in sync depending on just how similar the calibration densities are). This is to be expected since in such situations the implied number of links for both methods are approximately the same as the true number of links.

On the other hand, the more disparate the calibration densities for the two maximum entropy methods are, the more differences can be seen in the reconstruction performance as well as in the characteristics of the resulting networks. In such cases, ME Method 1 yields $TPR_{\mathcal{S}^{[\alpha]}}$ values and $PPV_{\mathcal{S}^{[\alpha]}}$ that are out of sync. As should be expected, when SN overestimates the true density (meaning that $c_{SN}^{[\alpha]} > c_{samp}^{[\alpha]}$), the $TPR_{\mathcal{S}^{[\alpha]}}$ values are larger than the $PPV_{\mathcal{S}^{[\alpha]}}$ values (the opposite also holds). Of course, the bigger the difference between $c_{SN}^{[\alpha]}$ and $c_{samp}^{[\alpha]}$, the more pronounced this difference becomes. In fact, the case with the largest difference between calibration densities, namely, commodity group 6100000 (the commodity group for telecommunication services), shows that the $TPR_{\mathcal{S}^{[\alpha]}}$ values, which were very high, and $PPV_{\mathcal{S}^{[\alpha]}}$ values, which were very low, of ME Method 1 almost exactly match those of the SN Method.

In addition to varying performances, the methods also revealed notable differences in the resulting estimated networks, which sheds light on the systematic biases arising in the SN Method with respect to the entropy-maximizing methods.

One repeated observation is that the out-degree distributions produced by the SN Method are more outward-shifted compared to those of the maximum entropy methods (see Figure 4.1(b)). Thus, despite the estimated power-law exponents being similar for all methods, the differences arising in the out-degree distributions of the network layers stem from the fact that the SN out-degree distribution (almost) always begins at a larger minimum out-degree and ends at a larger maximum out-degree. This systematic outward shift of the SN out-degree distribution is indicative of the incorrect link density produced by the SN Method.

Furthermore, a consistent result of the maximum entropy methods is that the emerging in-degree distributions always have a broad power-law regime with the size of the fitted exponent ranging between 2 and 3 (see Figure 4.1(a)). Despite the inability to test whether the estimates of the estimated power-law exponents correspond to their true values due to the lack of the complete real network data, this range of exponent values is what is observed universally across financial and economic networks studied in past literature, thus, rendering the resulting in-degree distributions a desirable and realistic property of the maximum entropy method

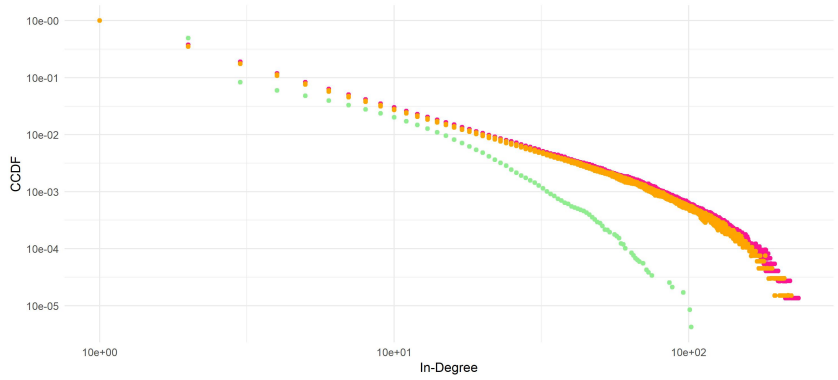
estimates. What is additionally special about this particular range of power-law exponents is that it indicates degree-heterogeneity (derived from the fact that when the network grows infinitely, the variance of the degree diverges). The heterogeneity of the degree is in fact the reason for many of the non-trivial features present in real-world networks. In contrast to ME Method 1 and ME Method 2, the SN method produces in-degree distributions with much larger estimated power-law exponents and finite variance when the network becomes increasingly large, ultimately indicating more degree-homogeneity than in the entropy-maximizing methods and in real-world networks.

In addition, the saturation for the in-strength as a function of in-degree for the SN Method also differs from those of the maximum entropy methods (see Figure 4.1(c)). For instance, the total in-strength is systematically achieved at much lower in-degrees in the SN networks and the rate at which the in-strength grows is not consistent as it tends to increase as degrees increase.

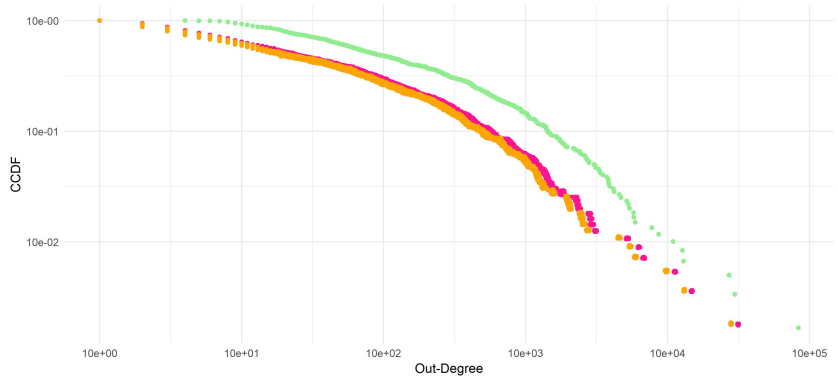
Regarding the other network statistics, one limitation of the maximum entropy methods with respect to the SN Method is their inability to capture the constraint that each user/supplier participating in the network should have at least one link (see Figure 4.1(d)). Indeed, in each realization, the ME Method 1 and ME Method 2 produce less links than the minimum expected degree. However, this is due to the probabilistic nature of maximum entropy methods meaning that the amount of zero-degree nodes varies between each realization and that certain nodes which were disconnected in one realization of the estimated network may be connected in another. Moreover, the SN Method always produces the network with the largest degree (even in situations where it does not estimate the most number of links). This explains its frequently lower value for the global clustering coefficient as well as shorter path lengths (see Rachkov (2020)).

In short, the identified biases arising in the SN Method with respect to the entropy-maximizing methods are: (1) the incorrect link density produced in the SN networks accounting for the systematic outward shift of its out-degree distributions; (2) the systematically high power-law exponents of the SN in-degree distributions and excessive degree-homogeneity; (3) the differing saturation of the in-strength as a function of in-degree; and (4) the systemically high value for the minimum expected degree. Note that the first three biases may be viewed as undesirable with respect to the maximum entropy methods, while the last mentioned bias may be viewed as desirable with respect to the maximum entropy methods.

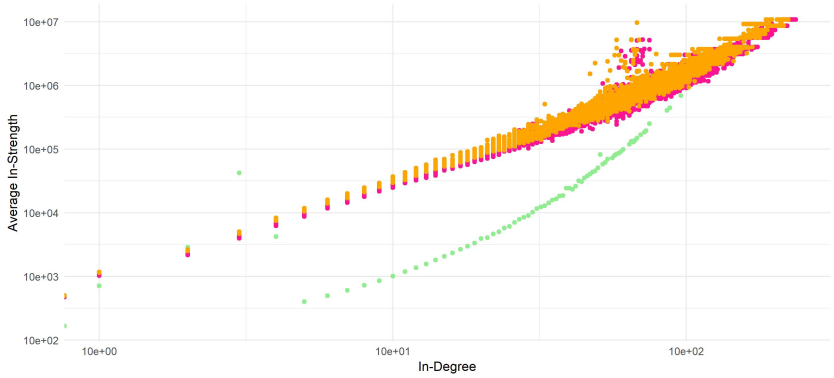
Figure 4.1 Example of Results - Commodity Group 3600000 (Water). Subplots (a), (b), and (c) depicted on the log-log scale. Green line depicts the SN Method, Orange depicts ME Method 1 and Magenta depicts ME Method 2.



(a) In-Degree Distribution



(b) Out-Degree Distribution



(c) Averaged In-Strength vs. In-Degree

4.2 Varying Density Performance Results

To test the effect of density on the performance of the maximum entropy method committing only to known information, a slightly modified version of ME Method 2 was applied to several commodity group subnetworks, the results of which are graphically displayed in Figure 4.2. Starting with Figure 4.2(a), it can be seen that, with the exception of a few outliers, it is generally the case that a higher true positive rate may be achieved at higher network densities. However, this is not surprising since a true positive rate equal to the density of the target network, i.e. $TPR = c$, indicates that the reconstruction method is only as good as a method which randomly assigns the expected number of links between nodes. The black line passing through the points

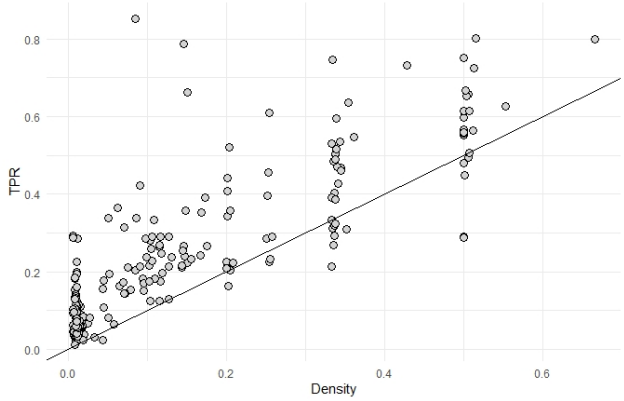
shows where the densities of networks would be exactly equal to the true positive rates. As one can see, from sparse to dense networks, there are always networks for which the true positive rate is much higher than the density. Nevertheless, there are also always some networks for which the true positive rate is only equal to the density or even less. The method can thus perform equally well or equally poorly at any density and there appears to exist some other factor that ameliorates or hinders its performance. Upon closer examination of a few cases where ME Method 2 performs either very well (i.e. $TPR \gg c$) or very poorly (i.e. $TPR < c$), it appears that several potential explanations can be discarded. These inlcude the total number of users and suppliers, $n_u + n_s$; the ratio of the number of users to suppliers, n_u/n_s; the shapes of the in- or out-strengths curves when plotted in ascending order; and the distribution of the in- or out-strengths.

Moving on to Figure 4.2(b), one can see that the reconstruction of sparser networks results in better specificity. This is logical because if there are extremely few links to be estimated as present within the network (relative to the number of possible links), then it is especially easy to estimate the absent links correctly. In the plot, specificity seems to decrease almost linearly as well as become gradually noisier with increasing density. Hence, specificity is more variable at higher densities.

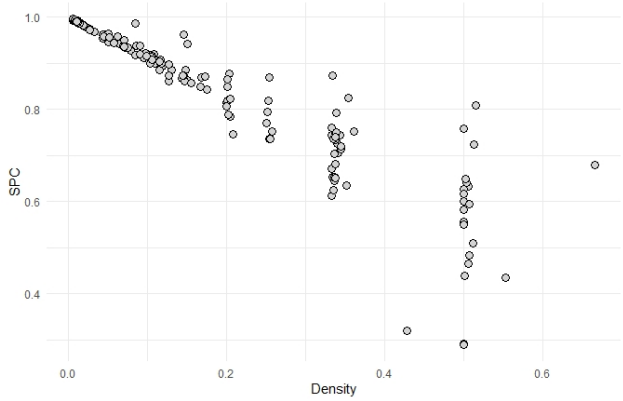
For the positive predicted values depicted in 4.2(c), the points are scattered in the same way as for the true positive rate in Figure 4.1(a). This is expected due to the way the method is calibrated. Recall that the unknown parameter $z^{[\alpha]}$ is computed by solving $L^{[\alpha]} = \langle L^{[\alpha]} \rangle$.

The last plot in Figure 4.2(d) displays the accuracy of the reconstructed networks of varying densities. The points are scattered in a similar fashion to those in Figure 4.2(b) with a few notable differences. Firstly, in the area with displaying sparser networks (where $c < 0.2$), the accuracy appears to decrease faster with increasing density compared to specificity. Moreover, there seems to be additional noisiness.

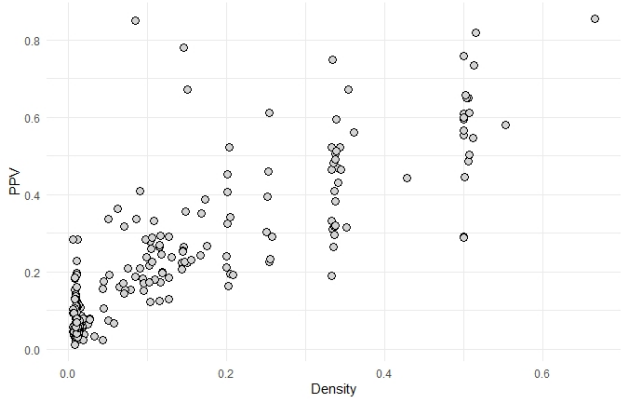
Figure 4.2 Performance Indicators as Functions of Density



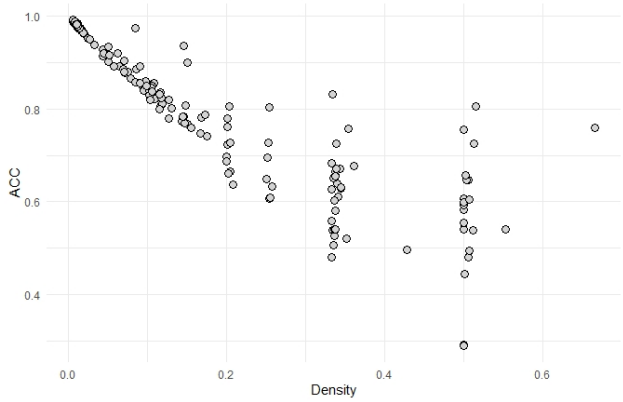
(a) True Positive Rate



(b) Specificity



(c) Positive Predictive Value



(d) Accuracy

5 Discussion

5.1 Business networks in the context of official statistics

Official business statistics as published by national statistical institutes (NSIs) such as Statistics Netherlands (SN) are currently mostly limited to descriptive statistics of the population: tables of various properties of businesses. While this certainly has its uses, many of the challenges faced by governance and policy require more in-depth understanding of the way in which the economy functions. In order to achieve this, it is necessary to have a better overview of the relationships between businesses. In this way it becomes possible to trace how businesses compete with each other, or whether they form part of a collaboration or even a production chain. These dependencies need to be mapped in order to understand endogeneous evolution of the population, as well as the response of the system of businesses to external economic shocks. This is important in the context of the current covid-19 pandemic with its economic impact. A system-level analysis is possibly even more important in the context of financial crises and exposure of the 'real' economy as well as the financial markets, to foreign debt, scarcity of rare resources, and climate change.

The Dutch economy is often described as an 'open economy'. This can mean several things depending on context. One aspect is regulatory: a company that is already established outside of the Netherlands and wishes to develop activities within this country has to deal with registration procedures and local laws and regulations. In any open economy, ideally these are transparent and integrated with EU and other international business practices so that they are a minimal hurdle. Another aspect, is the extent to which companies within the Netherlands buy and sell goods or services across borders. This also includes goods being shipped in through seaports which are subsequently exported via other means of transport. A substantial part of the Dutch GDP is generated through activities with a foreign component, and therefore also in this respect the Dutch economy can be said to be open. It is particularly this second aspect which is also relevant to this paper. This paper has focussed on business to business exchange of goods and services within the Netherlands. Given that there are strong connections to foreign business activity it would be beneficial to take this into account in a more granular fashion at a business and commodity level rather than treating all foreign trade as a single edge in the network. Relevant data is certainly available since for the purposes of payments of VAT for import and export a high quality register exists. The intention is to extend this work so that the Dutch business network can be embedded in an EU and worldwide network.

One of the aims of SN is to measure the system impact of those businesses that cease activities, which could be because of bankruptcy but there are other possibilities as well. If such a business is a central or crucial link in a production chain, the impact of ceasing operations is likely to be felt much more widely than if it is peripheral in the network or if its function is easily taken over by other businesses. This reasoning can be put on a more quantitative footing. For instance SN could not only count and publish the number of businesses that cease each month, but also a derivated quantity in which each of those businesses is weighted by their 'centrality' in the network. The centrality of a node in a network can be quantified in various ways. One appropriate way is to count for every node the proportion of paths of a given length where it is part of that path. If a node A has a high centrality in this sense, this means that if that node drops out of the network, many business B and C that connect through A can experience an interruption of the flow of goods or services. If there exist many alternative paths for two nodes

B and C in the network, that do not require A, ie. if the network redundancy is high, the impact on the system will be less dramatic. Also, if new connections in the network are established on relatively short timescales, the impact of any bankruptcy on the system might be mitigated. Effective government strategies to deal with the economic impact of potential surges in bankruptcies would benefit from detailed monitoring of indicators for the connectivity of the network and the extent to which the flow of goods or services is impeded as a consequence of crises. To the extent that a flourishing economy provides employment, and being in work tends to be positively associated with health and well being of the population, developing a suite of indicators to measure and monitor the state of the economic system has clear links to the sustainable development goals.

The need for efficient and unbiased network reconstruction methods thus becomes clear. In this paper, it was shown that the non-entropy-maximizing method developed by SN may indeed have a profound biasing effect, producing an incorrect resulting structure of the estimated networks. It is important to realize that these emerging biases may then affect one's ability to accurately capture indicators of connectivity or other relevant network statistics, which would ultimately hinder the strategies devised on the basis of these estimated networks. Truly harnessing the potential of network reconstructed methods therefore lies in the full understanding of the performance and behaviour of such methods. The paper at hand sets out to improve this understanding and highlighted the importance of contuining to do so in the future.

5.2 Further method extensions and challenges

In light of the above considerations, we can mention some further extensions of the method that are desirable as the object of future research.

It is important to firstly realize that the current research may in fact help in reducing the bias present in the method developed by SN (i.e. the SN Method). This is due to the fact that the maximum-entropy methods have usefully served as comparison material, ultimately indicating ways in which the existing assumptions underlying the SN Method's parameter values may be changed to mitigate biases in future network estimates. However, it is also important to note that the maximum-entropy methods offer a few significant advantages over the SN method, which perhaps warrants replacing the latter altogether. Although the SN method may appear to be more advantageous due to its deterministic nature (meaning that a single network is produced in each run of the reconstruction procedure), one can almost be certain that the true network is not the one which is being estimated and thus, the true network receives a likelihood of 0 rather than a desired 1. This is not the case for the maximum entropy methods, where the parameters are fit using maximum likelihood and therefore guarantee a positive likelihood for the true network. Furthermore, the maximum-entropy method offers a universal way of constructing networks that may be easily applied to any country. As a last remark, although the maximum-entropy methods are slightly more abstract and perhaps harder to grasp, this does not hinder the transparency and reproducibility of the methods and their outcomes.

The maximum-entropy method originates from the need of making the best (i.e. bias-minimising) use of the partial information available about an inter-firm network. In particular, the method used in this paper (ME Method 1) produces maximally random networks given aggregate data about the total in- and out-strength of each business, using as prior information a non-trivial relationship relating such flows to each node's expected number of

in-coming and out-going links respectively. While we implemented this exercise by focusing on the Dutch system and thereby reconstructing the (sector-specific) networks among firms in The Netherlands, in light of the economic openness mentioned above it would be important to allow for the reconstruction of links to foreign businesses as well. In particular, while in this paper the firm-specific in- and out-strengths were calculated as marginals of the local Dutch network, in general the pieces of information that are more easily accessible at the level of each individual firm are the overall sale and purchase volumes to/from the rest of the world. In terms of the network properties, these pieces of information correspond to the total in- and out-strengths of each node towards a set of nodes that is much bigger than the set of other nodes in the same country-specific network. While the current method allows for these other nodes to be treated in an aggregated fashion by formally appending one additional 'rest of the world' supernode to the domestic inter-firm network, it would be desirable for the method to combine information coming from, say, statistical offices of several countries to partly resolve the supernode into the individual firms located in these other countries.

Similarly, it would be important for a reconstruction method to combine the local firm-specific marginals with global information coming from input-output tables at both domestic and multi-regional levels. This means that the connection probability should contain an expanded set of parameters that allow for the inclusion of sectoral information obtained from 'macroscopic' input-output tables reporting flows among economic sectors. This would lead to a framework where either the in- and out-strength of each firm are disaggregated into sectoral components and separately taken as input information, or the density parameter is made sector-dependent.

Besides expanding the model with macroscopic information, it would of course be desirable to allow for finer-grained, 'microscopic' information about actual transactions between individual firms. Inter-firm transaction data are available e.g. at private banks in the form of payment flows among the accounts of firms that are customers of the same bank. Clearly, these data are protected by confidentiality and cannot be accessed publicly. Still, it is conceivable that, for the mere purpose of parameter estimation, statistical offices and universities may cooperate with private banks in order to calibrate network reconstruction models on fully anonymized transaction data. Moreover, in order to safely integrate information available at multiple banks, one may think of an approach where each bank provides only the (again anonymized) margins of the inter-firm transaction network constructed from their own data. In principle, collecting the margins from multiple banks should respect confidentiality of the original information, while at the same time providing an integrated account of the local firm-specific properties that are crucial for a good performance of the network reconstruction method.

Finally, from a technical point of view, a reconstruction method that is capable of integrating information coming from multiple scales and levels of resolution as advocated above should have good properties in terms of statistical estimation. In particular, the estimated probability of connection between two nodes in the same subnetwork, for which the marginal information has already been used, should be robust to e.g. a subsequent estimation when more nodes are added to the system, each carrying its additional piece of marginal information. The desirable property of robustness of the statistical estimation of the parameters of a network model with respect to the inclusion of additional nodes, or in general to a change in the scale of resolution of the network, goes under the name of *projectivity* in statistics (Shalizi and Rinaldo (2013); Krioukov and Ostilli (2013)). Projectivity ensures that, if one regards a network as a subgraph of a larger network, the edges of the subgraph are sampled from the same distribution as the those of the original smaller network. The absence of this requirement is clearly undesirable.

To summarize, we believe that an important challenge for future research in the field of network reconstruction for official statistics is the development of generalized multi-country and multi-scale methods that allow for the representation of a domestic inter-firm network as a portion of a larger network where also links to foreign businesses can be treated consistently and where the empirical knowledge can be augmented by including both 'macroscopic' (possibly multi-regional) input-output tables and 'microscopic' firm-to-firm transaction data. All this should be implemented within a projective network model. Ideally, such methodological extension should go hand in hand with more detailed data made available by statistical offices of multiple countries, disaggregating each firm's in- and out-flows to each country-sector pair in a harmonized and consistent way. Similarly, it would be desirable to explore privacy-compliant ways to share and integrate suitably marginalized firm-specific information obtained from transaction data available at private banks, in such a way that the parameters of the reconstruction model can be estimated much better, while at the same time preserving the confidentiality of sensitive information.

References

Anand, K., I. van Lelyveld, Ádám Banai, S. Friedrich, R. Garratt, G. Hałaj, J. Fique, I. Hansen, S. M. Jaramillo, H. Lee, J. L. Molina-Borboa, S. Nobili, S. Rajan, D. Salakhova, T. C. Silva, L. Silvestri, and S. R. S. de Souza (2017). The missing links: A global study on uncovering financial network structures from partial data. ESRB Working Paper Series 51, European Systemic Risk Board (ESRB), European System of Financial Supervision.

Baldi, P. (2002, January). A computational theory of surprise. In *Information, Coding and Mathematics*. Springer.

Barabási, A. L. and M. Pósfai (2016). *Network science*. Cambridge: Cambridge University Press.

Bliss, C. A., C. M. Danforth, and P. S. Dodds (2014, October). Estimation of global network statistics from incomplete data. *PLoS ONE 9*(10).

Borgatti, S., K. Carley, and D. Krackhardt (2006, May). On the robustness of centrality measures under conditions of imperfect data. *Social Networks 28*.

Caldarelli, G., A. Capocci, P. D. L. Rios, and M. Munoz (2002). Scale-Free Networks from Varying Vertex Intrinsic Fitness. *Physical Review Letter 89*(25).

Chen, L., A. Karbasi, and F. Crawford (2016, October). Estimating the size of a large network and its communities from a random sample. *Advances in neural information processing systems 29*.

Cimini, G., T. Squartini, D. Garlaschelli, and A. Gabrielli (2015, October). Systemic risk analysis on reconstructed economic and financial networks. *Scientific Reports 5*.

De Martino, A. and D. De Martino (2018, April). An introduction to the maximum entropy approach and its application to inference problems in biology. *Heliyon 4*.

Garlaschelli, D. and M. I. Loffredo (2004, October). Fitness-dependent topological properties of the world trade web. *Physical Review Letters 93*(18).

Hooijmaaijers, S. and G. Buiten (2019, April). A methodology for estimating the dutch interfirm trade network, including a breakdown by commodity.

Hsieh, C., S. Ko, J. Kovarik, and T. Logan (2018). Non-randomly sampled networks: Biases and corrections. Technical report, National Bureau of Economic Research, Inc.

Huisman, M. (2009). Imputation of missing network data: Some simple procedures. *Journal of Social Structure 10*(1).

Jaynes, E. T. (1957, May). Information theory and statistical mechanics. *Phys. Rev. 106*(4).

Kim, M. and J. Leskovec (2011, April). The network completion problem: Inferring missing nodes and edges in networks. In *Proceedings of the 11th SIAM International Conference on Data Mining, SDM 2011*.

Krioukov, D. and M. Ostilli (2013). Duality between equilibrium and growing networks. *Phys. Rev. E 88*(2).

Lebacher, M., S. Cook, N. Klein, and G. Kauermann (2019, September). In search of lost edges: A case study on reconstructing financial networks.

Lerner, J., D. Wagner, and K. Zweig (2009, January). *Algorithmics of large and complex networks. Design, analysis, and simulation*, Volume 5515. Springer.

Mazzarisi, P. and F. Lillo (2017, January). *Methods for Reconstructing Interbank Networks from Limited Information: A Comparison*, pp. 201–215. Springer.

Morais, M. (2018). Statistical modeling and analysis of neural data: Information theory and maximum entropy. Lecture Notes.

Parisi, F., T. Squartini, and D. Garlaschelli (2018). A faster horse on a safer trail: generalized inference for the efficient reconstruction of weighted networks.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rachkov, A. (2020). Bias in non-entropy-maximizing network reconstruction methods.

Ramadiah, A., F. Caccioli, and D. Fricke (2017, January). Reconstructing and stress testing credit networks. Technical report, European Systemic Risk Board (ESRB), European System of Financial Supervision.

Robinson, D. (2008, 12). Entropy and uncertainty. *Entropy: International and Interdisciplinary Journal of Entropy and Information Studies 10*.

Shalizi, C. and A. Rinaldo (2013). Consistency under sampling of exponential random graph models. *Ann. Stat. 41*(2).

Smith, J. and J. Moody (2013, October). Structural effects of network sampling coverage i: Nodes missing at random. *Social Networks 35*.

Smith, J., J. Moody, and J. Morgan (2017, January). Network sampling coverage ii: The effect of non-random missing data on network measurement. *Social Networks 48*.

Squartini, T., G. Caldarelli, G. Cimini, A. Gabrielli, and D. Garlaschelli (2018, October). Reconstruction methods for networks: The case of economic and financial systems. *Physics Reports 757*.

Squartini, T., G. Cimini, A. Gabrielli, and D. Garlaschelli (2017, January). Network reconstruction via density sampling. *Applied Network Science 2*(1).

Squartini, T., R. Mastrandrea, and D. Garlaschelli (2015, February). Unbiased sampling of network ensembles. *New Journal of Physics 17*.

van Steen, M. (2010, January). *Graph Theory and Complex Networks: An Introduction*. Steen.

Watamabe, H., H. Takayasu, and M. Takayasu (2013, February). Relations between allometric scalings and fluctuations in complex systems: The case of japanese firms. *Physica A Statistical and Theoretical Physics 392*.

Wilson, R. J. (1996). *Introduction to Graph Theory* (Fourth ed.). John Wiley & Sons, Inc.

Zhang, Y., E. Kolaczyk, and B. Spencer (2015, May). Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *The Annals of Applied Statistics 9*.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information