Discussion Paper

Stochastically reconstructing business production networks using maximum entropy

J.O. Kayzel
F.P. Pijpers

June 2023

Contents

Analysing business production networks is a pioneering field for national statistical institutes. A combination of available data and reconstruction techniques makes it possible to map out business relations between firms within a system. Given the limits of the information that is available for the Netherlands, it is hard to map out these networks. The application of network reconstruction methods are the main focus of this paper. These reconstruction techniques are applied to the Dutch interfirm trade-network, i.e. the commodity trading activity between firms in the Netherlands. The focus is networks from a static point of view i.e. looking at the configuration of the networks at a certain point in time. It is, however, also interesting to look at the time-dimension of such networks. For these time-dependent networks certain dynamical indicators can be investigated. This way one can see the effects of changes to the network when links appear/disappear or when weights get adjusted.

1 Introduction

Analysing business production networks is a pioneering field for national statistical institutes (NSIs), universities and other organisations. A combination of available data and reconstruction techniques makes it possible to map out business relations between firms within a system. Such a system consists of many actors and their interactions, often resulting in an intricate structure, thus being referred to as a complex system. Viewing an economic system as complex, allows one to formulate an alternative that sits somewhere between the general equilibrium theory of neoclassical supply & demand economics and the increasingly data-driven fields within econometrics.

It is in the interest of Statistics Netherlands(SN) to have detailed knowledge on such economic systems. It allows for the investigation of emergent behaviours and structure in the system. These emergent patterns are not part of the system's data-input or model assumptions, but is inherently and consistently generated by the structure of the network, thus giving new insights in the networks' features. The direct application of such knowledge is the ability to asses risks or predict the effect of shocks within the system.

A well known example for such a risk is the 2007/2008 financial crisis. It reached a tipping point when investment bank Lehman Brothers declared bankruptcy, causing a cascade effect on the banking sector. The recent (February 2022) supply-chain issues are thought to be (partially) caused by a few, but very vital bottlenecks, like the shortage of a single computer-chip component or the blocking of one shipping strait. Viewing large-scale economies as a network of relationships between firms and sectors, the configuration of the network is very detailed, allowing zooming in at the level of individual firms. As such, it can help locate such vulnerabilities within a system. Also, the impact of long-term potential threats, such as e.g. climate change, can only be understood by tracing through their impact on individual firms and investigate how central, in network terms, the role of those firms is. A small impact on a firm with a very central role in the network may well have far more significant consequences, than a big impact on a firm that is really peripheral in terms of connections to other firms.

It is not only shocks that originate from abroad and then propagate through the internal Dutch economic system that are of interest. Also from within the Dutch internal market system itself instabilities might arise, or cyclical behaviour, which has significant impact on aggregated properties such as economic growth, inflation, or innovation and productivity of labour. These latter quantities are monitored as part of the National Accounts of the Netherlands because of their macro-economic importance, but they are emergent properties: generated by the conglomerate of many individual transactions and exchanges of goods and services between individual firms. One could regard something like business innovation as a 'contagious' property so that innovative businesses may have a tendency to form more tightly connected subclusters within the overall network. Identifying such communities, for instance to better target packages of stimulation programmes is of direct policy relevance. A better understanding of the economy, and therefore improvements in the ability for forecasting by the Dutch Bureau of Economic Policy Analysis (CPB), relies on understanding the mechanisms that produce these emergent properties.

In neoclassical-economics the existence of rational agents is assumed, who make informed decisions about actions in an economic system. It is thought that each actor in a system will make the same rational choice regarding problems, and thus the system moves towards; or rotates around a general equilibrium. While economists are generally aware of this unrealistic expectation, it allows making relatively simple models to explain economic systems on a macro scale without the requirement of microscopic details. On the other hand, there is also a trend in using data-driven models where the conclusions are drawn based on (large) data sets attained from empirical studies. However, these methods are often limited to micro-economic or small-scale macro-economic research[1)]. In complexity economics there is an effort to strike a balance between simplistic macro models and more data-driven micro scale models. The idea is to make a model that captures the macro scale picture of a system that, at the same time, also allows zooming in on the more individual micro scale level. This can be achieved through a combination of available micro-data, and observed macro properties (Arthur, 2021).

It is often hard to map out these networks when only limited information is available. The application of such network reconstruction methods are the main focus of this paper. These reconstruction techniques will be applied to the Dutch interfirm trade-network, i.e. the commodity trading activity between firms in the Netherlands. The focus will be networks from a static point of view i.e. looking at the configuration of the networks at a certain point in time. It is, however, also interesting to look at the time-dimension of such networks. For these time-dependent networks certain dynamical indicators can be investigated. In this way one can see the effects of changes to the network when links appear/disappear or when weights get adjusted. When analysing risks it would be of interest, for instance, to analyse changes to the overall network topology when a link is removed. If, as a result, the structure changes significantly it might signify a structural risk in the system. It is not the goal of this paper to research dynamical indicators, but it will serve as a motivation for the reconstruction itself. The hope of this particular research is to accurately and confidently 'predict' the configuration of the network based on partial information. Thus providing a strong foundation for any subsequent investigation into risk-assessment or other (dynamical) indicators of interest (Squartini et al., 2018). This research builds on existing methods that have mostly been tested out on networks which are slightly denser in terms of the links between firms, and where also a little more is known about the links, such as interbank lending. The extension to the Dutch (internal) trade network was first attempted in Rachkov et al. (2021) and is extended more systematically here. In particular the aim is to perform a sensitivity analysis by various techniques, and making use of the multilevel nature of the data, where firms can be considered as (predominantly) operating within one or very few 'layers' of the network, where each layer is a sector. The approach is probabilistic rather than deterministic, such as for example IPF/RAS, as is detailed in the next section because this makes more objective use of subsamples where more details about the network are known. Just as in the previous study, trade with firms that are registered outside of the Netherlands are excluded from the system. While for the Dutch economy foreign trade is certainly important, its full reconstruction is beyond the scope of feasibility for this research, whereas the structure of interfirm trade activity within the Netherlands only is of interest by itself.

1) For example: The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

1.1 Deterministic and probabilistic methods

For trade-networks several layers of aggregation can be considered in the reconstruction procedure. For instance, using the GDP of countries, an international trade-web could be mapped out. On a lower level, it is possible to look at firms in certain industries (e.g. agriculture, construction, finance etc) and the contribution of each to the total quantity or value of goods or services and make an intercommodity-group network. Or, another layer below, the trade between firms in a commodity group (e.g. wheat, architecture, banking etc). It is the latter layer that will be the focus point of this paper, but the relevant reconstruction procedures can be carried out on any layer, if the necessary information is available.

In the gathering of information to map out the commodity network, often privacy and disclosure issues are at play, resulting in knowing only part of the network configuration. In some countries, such as for instance Japan or Belgium, tax regulations are such that for every transaction between firms an electronic record is kept of the associated VAT owed. Aggregates of such records are far more detailed than can be found in annual business reports. In the Netherlands, such detailed information is not required and no such register is available, so it is necessary to rely on a sample of businesses, which is somewhat biased since the source available contains only business with large to very large turnovers, and which are surveyed by a business (lending) risk assessor. The goal is to find suitable models and estimators to reconstruct the networks using the limited information there is available. For the interfirm trade case, for each firm the total ingoing and outgoing trade-volume is known because SN has records of annual revenue and profits for each firm. Moreover, on part of the network the amount of trading from a sample of the total network is known. The samples on these activities are obtained by SN via an external company. In the current model employed by SN, the network is constructed based on relationships between the known information and the unknown pieces. The relationships are empirically substantiated by having access to trade-networks from other countries where this detailed information is available. It then rank-orders each firm from 'strong' to 'weak' and predeterminedly distributes the available links in the system according to the firms strength-level.

The downside of this deterministic method is that it only produces one answer, and by the nature of model-misspecification almost surely that answer will be wrong. Any mistake in the model assumptions, no matter how small, will result in a wrong output. It is therefore in the interest of SN to try and use a more probabilistic method to reconstruct networks. The main idea is that by using a probabilistic model, the output could be an entire ensemble of possible configurations that adhere to the known information. As such, it could likely still be the case that this ensemble does not contain the true configuration. However, by taking in a probabilistic viewpoint, statistical tools can be used to evaluate predictive power of the reconstructed network and ask common statistical questions. Is the answer approximately correct? What is the chance of the ensemble containing the correct answer? Is the model consistent?

To use a probabilistic model, a probability distribution needs to be found that assigns to each pair of firms the probability of them trading with each other. This is opposed to the algorithmic distribution of links in the deterministic method. To gain knowledge on a system based on partial information, one often wishes to maximise the underlying

likelihood to arrive at an estimator for the parameters in the model distribution. Using a more general likelihood-maximisation method here is not desirable, as networks are often too heterogeneous to be inferred from the available information. Thus, the likelihood is constrained w.r.t. the known information. In SN's case, only the trade volumes or in-and out-strengths per node/firm are known and not the degree of each node (the number of edges, or relations to other firms, it has). The constraints used in the maximisation process (based on what is known) are often not sufficient statistics for the likelihood-function of the network as a whole (Parisi et al., 2020). Thus, leaning too much on the density of the partial network results in bias for the estimator. To find the likelihood, an entropy framework (based on information theory) can be used to reconstruct networks constrained to known information. Shannon Entropy can be viewed as a measure for (average) uncertainty. The more uncertainty there is about the configuration of the network, the more information is needed for its reconstruction. If the Shannon entropy of the system is maximised, conditional only on the available information, one becomes maximally non-committal on unknown information. This will result in the least biased estimator for such a system. Lagrange optimisation is utilised to derive a probability distribution/density function for the desired network. If the constraints of the maximisation are only macroscopic moment information, the subsequent density that arises from this method is from an exponential family.Thus, a theory using Exponential Random Graphs(ERG) will be adopted where the existence and weights of links is described via a distribution from an exponential family.

From this entropy maximisation theory, multiple methods for reconstructing the network can be described. While some of these methods are deterministic in nature, others, like the entropy-maximisation method using average constraints, are probabilistic. Various methods can be used in both frameworks (Squartini et al., 2018).

1.2 Research Outline

The methods featured in this paper are extensively researched, discussed and applied in (Rachkov et al., 2021; Squartini et al., 2018; Parisi et al., 2020; Cimini et al., 2015) and many other papers, mainly involving maximum entropy configuration methods. There, they are often applied to reconstruct the World Trade Web (WTW) or E-mid banking system or similar systems. For the research reported here, these methods are applied to the Dutch interfirm-trade network, which is different for several reasons. Mainly the amount of known information is severely limited, and the number of acting nodes in the interfirm-system is significantly higher than the WTW and E-mid systems. There are various details that are different within each system that should be taken into consideration.

The topic of this discussion paper is a continuation of earlier research done by Andrea Rachkov at SN (see (Rachkov et al., 2021)). In this earlier work on the topic of interfirm trade-networks, the deterministic SN method was directly compared to a maximum entropy approach called the Fitness Model. Aside from a direct comparison, the performance of first order, i.e. node/link specific, statistical indicators like accuracy and the positive predictive value were evaluated for this Fitness Model.

The advantage of the fitness model is that it allows the production an ensemble of configurations while not totally abandoning the known relationships used in the

deterministic method. Moreover, in the setting of economic networks, it is assumed that the network features a good-get-richer phenomenon. This expresses an expectation that firms with higher in/out-strength (e.g. firms with higher/lower revenue or turnover) are thought to be more attractive to trade with for other firms. This will result in the forming of so called k-stars (k referring to the degree) and clustering, i.e. the occurrence of triadic motifs, in the resulting network (Kolaczyk, 2007). It has been shown that incorporating such a phenomenon into the fitness model, will allow it to produce these higher order attributes better than the deterministic method is able to (Rachkov et al., 2021; Cimini et al., 2015).

The first goal of this paper is explaining and expanding the fitness model. therefore, section 2 presents a detailed look at the derivation of the exponential random graph distribution that is produced from maximising entropy which is itself a concept from information theory. Moreover, this ERG-framework leads to the class of Configuration models, to which the fitness model belongs. The logic behind the fitness model is explained and why this specific model is chosen. Previous work is expanded upon by adding weights (i.e. the trade-volume per link) to the reconstruction method. This leads to a probabilistic method called the Conditional Reconstruction Method(CReM) that accurately and consistently assigns weights in a probabilistic way.

In section 3 the choice of models is discussed, as well as how to asses their performance. Finding out whether the reconstruction methods actually produces realistic results is difficult, as SN has no exact knowledge of the true configuration of the network. In order to asses the performance of a model it can be compared with other models as done inRachkov et al. (2021). It could also be compared to the expectation of the model and its model-specification(as seen in Squartini et al. (2018); Parisi et al. (2020); Cimini et al. (2015)). Where possible results are compared with the theoretically expected or known values of the network, but often these cause issues in the setting of large networks. Therefore, a validation scheme is employed to asses the performance of the method and determine its consistency and robustness.

Section 4 evaluates the ensemble outputs of the reconstruction procedure using these methods. Several attributes of the resulting reconstructions are investigated, where often the behaviour of the ensemble averages is investigated. The focus is on higher order attributes, concerning not the nodes in the network itself but the patterns that arise in the direct neighbourhood of nodes. A bit more information on first order attributes can be found in appendix B, but a thorough investigation into their performance can be found in Rachkov (2020).

To investigate sampling bias present in the Fitness model, a resampling scheme is utilised to study the possibility of alleviating potential sampling bias even more. When samples are too small to arrive at asymptotic normal confidence intervals, a bootstrap-scheme can be used to arrive at suitable confidence intervals for most performance measures. Here, a slightly altered studentised bootstrap-scheme is employed to construct confidence intervals with small samples. Lastly, the results and possible future research are discussed.

2 Theoretical background

Entropy is a concept from thermodynamics and from information theory, which is used to quantify the information contained within certain observations by considering (i.e. counting) bits of information. Entropy maximization is a well-known fitting technique that can also be used to reconstruct networks from limited available data[2)]. To actually derive a probability distribution from the entropy framework, a Lagrangian optimisation is used on the network. In network theory the matrix representation of graphs is often used as a way both to make large data sets less cluttered and to be able to use algebra on them. Then the Lagrangian optimisation is used to maximise the entropy constrained w.r.t. known information, that gives us an Exponential Random Graph (ERG) distribution, which will be the framework for the models discussed here. Specifically, Shannon entropy is introduced. Maximising it will make the analysis maximally non-committed to the unknown part of the network, thus alleviating potential bias.

2.1 Network Representation

A network G is seen as a set of vertices V and edges E between vertices. However, in the context of networks they are often called nodes and links respectively. In order to effectively represent large networks of thousands or more firms, networks are conveniently summarised by their matrix representation. The Adjacency matrix of a graph G with N vertices is denoted as $A(G)$. It is a square matrix with entries a_{ij}, $i, j \in \{1, ..., N\}$ such that

$$a_{ij} = \begin{cases} 1 & \text{if } (v_i, v_j) \in E \\ 0 & \text{else.} \end{cases} \tag{1}$$

The number of direct links a node has is called the degree of a node. The set of all degrees is called the degree sequence. The (link)-density d of a graph (without self-loops) is given by the fraction of edges in a graph w.r.t. the total number of possible edges $N(N-1)$ in a graph. If $|V| = N$ and $|E| = L$ it is given by

$$d = \frac{L}{N(N-1)}. \tag{2}$$

This is assuming the presence of directions, otherwise the number of possible edges is $\frac{1}{2}N(N-1)$ so that the righthand side of (2) would need to be multiplied by a factor of 2.

The adjacency matrix can be enriched with weights $w_{ij} \geq 0$ such that one can assign a certain weight $w_{ij} > 0$ for each edge (v_i, v_j) in the graph, and 0 otherwise. The resulting matrix will be denoted as W and its weighted degree is called its strength. Each entry in the matrix represents a (weighted) edge between two nodes. The set of all the nodes' strengths is called the strength sequence.

2) Detailed and more general descriptions can also be found in Squartini et al. (2018); Kolaczyk (2007); Squartini and Garlaschelli (2017)

2.2 Entropy maximisation Framework

In information theory, Shannon Entropy is used to quantify how many 'bits' of information are obtained from an observation. The main idea is two-fold. First, it is used to measure how flat a probability distribution is. The closer the distribution is to uniform the higher the entropy will be. It can thus be seen as the Kullback-Leibler divergence from the uniform distribution. Secondly, it can quantify how many possibilities there are in general for a given outcome.

The idea is then to have a measure on the variability of the system of interest. For instance, given that a portion of a network is known, how does one expect the rest of the network to be configured? If the network has a low measure of variability, it is expected that the total network resembles the known portion. Conversely, for a high variability it may be expected that the rest of the network is different from the known portion. Without extra assumptions the unknown portion is thus considered to be as uniformly distributed as possible, to translate the notion of being uninformed. So the question is, how much information is gained from observing a part of the network?

An obvious candidate for variability might be the variance. However, the variance measures the proportional spread of outcomes, and as such the variability measured by the variance is heavily influenced by the relative proportion of the different events in the sample space. With network reconstruction (and specifically link-incidence reconstruction) the quantity of interest is the variability in a topological sense. In other words, the quantity should say something about the number of different possible configurations, where no configuration is 'more different' than any other. For this purpose, a so called information measure is introduced.

Given a random variable X and a realisation x, a measure of uncertainty I, or information needed to describe a system, is given as

$$I_X(x) = -\log\left(P(\{X = x\}\right) \tag{3}$$

Note that $I_X(x)$ is 0 if the event has probability 1. This implies that, when there is complete certainty of an outcome there is 0 uncertainty, i.e no new information is gained by the outcome. On the other hand, for an event that has probability approaching 0, the uncertainty will tend to infinity so such an outcome reveals a lot about the system. The natural logarithm functions as a base for the quantity of information. This information measure can be averaged out over all outcomes to arrive at a definition for Shannon Entropy:

$$\begin{aligned} S(X) &= \sum_{\{X=x\}\in\Omega} P_X(x) I_X(x) \\ &= -\sum_{\{X=x\}\in\Omega} P_X(x) \log P_X(x) \end{aligned} \tag{4}$$

The notation is simplified by omitting subscript X, writing $I(x) = -\log p(x)$ and $S = -\sum_x p(x) \log p(x)$. As the input of the Shannon-entropy is often obvious from context (the system of interest) it is simply denoted as S.

A problem regularly encountered with real-world networks, or at least those relevant in this paper, is that they are sparse with high-degree clusters. The implication is that they

feature (fully) connected subgroups of vertices of high degree nodes, and have low-degree 'out-lying' nodes elsewhere. With standard sampling techniques the resulting reconstruction will possibly be heavily biased. This is where Shannon entropy comes into play.

Intuitively, Shannon entropy can be seen as the expected degree of surprise. If a given event is totally expected, that is has likelihood $= 1$, its occurrence provides no new information: $p(x) = 1 \Rightarrow S = 0$. If, on the other hand, that event was very unexpected, its occurrence will give a lot of information: $p(x) \downarrow 0 \Rightarrow S \uparrow \infty$. Entropy can thus also be viewed as a measure of how confident one is to predict new outcomes. Maximising entropy, constrained by what is known, makes one maximally non-committed to what is unknown. I.e. the known (observed) portion of the system gives as little information as possible. The belief is that when using this idea, the estimates and predictions contain the smallest amount of bias. The reason one would want to adopt this maximisation has to do with the lack of projectivity in the system of interest (Clauset et al., 2009). When using partial information to reconstruct the whole, a lack of projectivity means that this partial information tells us nothing about the configuration of the rest of the system. Thus, using the partial information as a predictor for the rest of the system can be seen as inducing bias. Instead, the partial information is used as constraints on the systems configuration whilst maximising its entropy given these constraints. This takes the heterogeneity of the system into account and thus does not project a selective part on the whole.

The known constraints are often, in the reconstruction framework, particular aggregates of information on the system. For this framework the known constraints are defined as the available moment information. A set of moment functions $f_m(x)$ are used to derive one or more constraints

$$\mathbb{E}f_m(X) = \sum_x p(x) f_m(x), \text{ for } m = 1, ..., M. \tag{5}$$

To maximise entropy constrained w.r.t. the available moment information, a Lagrange optimisation problem, of which a detailed derivation and explanation can be found in Kelly and Yudovina (2014), is defined:

$$\begin{aligned} \text{maximise} \quad S &= -\sum_x p(x) \log p(x) \\ \text{subject to} \quad \mathbb{E}f_m(X) &= \sum_x p(x) f_m(x), \text{ for } m = 1, ..., M \end{aligned} \tag{6}$$

For normalisation purposes, the constraint is added that $f_0 = 1$ and $\mathbb{E}f_0(X) = 1$. This will ensure that the result is in fact a properly normalised probability distribution. This setup produces the Lagrangian

$$\begin{aligned} \mathcal{L}(P) &= -\sum_x p(x) \log p(x) - \lambda_0 \left(\sum_x p(x) - 1 \right) \\ &\quad - \sum_{m=1}^{M} \lambda_m \left(\sum_x p(x) f_m(x) - \mathbb{E}f_m(X) \right). \end{aligned} \tag{7}$$

with λ_m the Lagrange-multiplier corresponding to moment-information/constraint f_m.

The next step is to solve $\frac{dL(P)}{dp(x)} = 0$ w.r.t. each marginal probability $p(x)$. That means:

$$\begin{aligned} \frac{d\mathcal{L}(p(x))}{dp(x)} &= -(\log p(x) + 1) - \lambda_0 - \sum_{m=1}^{M} \lambda_m f_m(x) = 0 \\ p(x) &= \frac{e^{-\sum_{m=1}^{M} \lambda_m f_m(x)}}{e^{\lambda_0 + 1}} \end{aligned} \tag{8}$$

The normalisation condition, from dealing with probabilty distributions, as a constraint means that it must hold that $\sum_x p(x) = 1$, which in fact sets the value of the Lagrange multiplier λ_0. Since e^{λ_0+1} must enforce this constraint,

$$e^{\lambda_0+1} = \sum_x e^{-\sum_{m=1}^{M} \lambda_m f_m(x)}. \tag{9}$$

The resulting solution produces the probability distribution defined by:

$$p(x) = \frac{e^{-\sum_{m=1}^{M} \lambda_m f_m(x)}}{\sum_x e^{-\sum_{m=1}^{M} \lambda_m f_m(x)}} \tag{10}$$

where λ_m denotes the Lagrange multiplier corresponding to the m-th moment information constraint.

This distribution has the form of an Exponential random graph(ERG) distribution with (general) distribution function

$$P(X = x|\theta) = \frac{e^{\theta^T T(x)}}{c(\theta)} \tag{11}$$

where θ^T is a tuning parameter, c a normalising constant and T some statistic.

This probability distribution function is the main workhorse for the reconstruction methods discussed here. The exponent of the ERG is the Hamiltonian $H(G|\lambda)$ which represents the chosen constraints with Lagrange-multipliers λ, thus giving:

$$p(G|\lambda) = \frac{e^{-H(G|\lambda)}}{c(\lambda)} \tag{12}$$

where $c(\lambda) = \sum_G e^{-H(G|\lambda)}$.

Note that, when the Shannon entropy is maximised without constraints, the resulting solution is the uniform distribution $p(G) = 1/\{\#\text{ of possible configurations of G}\}$. This is in line with a property of the Shannon entropy, that it attains its maximum when unconstrained. Moreover, this reinforces the view of the Shannon entropy as the Kullback-Leibler divergence from the uniform distribution. Kullback-Leibler divergence (Kullback and Leibler, 1951) is an information-based measure of disparity among probability distributions. Given distributions P and Q defined over X, with Q absolutely continuous with respect to P, the Kullback-Leibler divergence of Q from P is the P-expectation of $-\log_2\{P/Q\}$, or:

$$D_{KL} = \int_X -\log_2 \left[\frac{Q(x)}{P(x)}\right] \mathrm{d}P \tag{13}$$

Lastly, the constraints are given as $\{\mathbb{E}f_m(\mathcal{G})\}_{m=1}^{M}$ for $\lambda = (\lambda_1, ..., \lambda_m)$ in an optimal way[3)].

3) Where $\mathcal{G}$ can be seen as the random variable with possible realisation (configuration) G.

By the Maximum Likelihood Estimator(MLE)-method the suitable parameters are set. Recall that

$$\mathbb{E}f_m(\mathcal{G}) = \sum_G P(G|\lambda) f_m(G). \tag{14}$$

With the MLE, the desired result is acquired via the log-likelihood. So for each m and observed G^*:

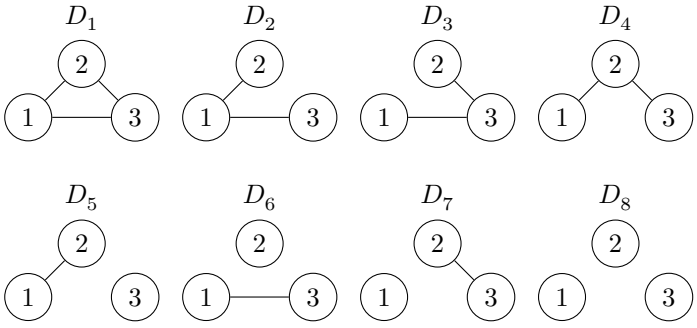
$$f_m(G^*) = \frac{\sum_G f_m(G) e^{-\sum_{m=1}^{m} \lambda_m f_m(G)}}{c(\lambda)} = \mathbb{E}f_m(\mathcal{G}) \tag{15}$$

The nice result here is that, as can be seen here, the most likely Lagrange-multipliers w.r.t. a given configuration are in full agreement with the desired constraints arising from the entropy optimisation. These constraints are encoded within the Hamiltonian. Often they are taken as the (known) degree and strength sequences. Given a network G represented by an $N \times N$ adjacency matrix A, the in/out-degrees of a node i are denoted as k_i^{in} resp. k_i^{out} and the in/out-strength of a node i as s_i^{in} resp. s_i^{out} . Defined as

$$\begin{aligned} k_i^{in} &= \sum_{j=1}^{N} a_{ji}, \qquad s_i^{in} = \sum_{j=1}^{N} w_{ji} \\ k_i^{out} &= \sum_{j=1}^{N} a_{ij}, \qquad s_i^{out} = \sum_{j=1}^{N} w_{ij} \end{aligned} \tag{16}$$

Notice that $\mathbb{E}a_{ij} = p_{ij}$ with p_{ij} being the probability there exists a link between node i and j. Via the adjacency matrix, also a link-incidence probability matrix P with entries p_{ij} is obtained.

Example 1. A small example can give an intuitive understanding of the method employed. Consider a network with nodes $i = 1, 2, 3$. This system has the following possible network configurations D_i:



If there is no information on the system, the best possible guess would be a uniform random guess. The probability of getting the right configuration is then $\frac{1}{8}$. The corresponding entropy is

$$\begin{aligned} S &= -\sum_{i=1}^{8} p(D_i) \log D_i \\ &= 8 \times \left(-\frac{1}{8} \log \frac{1}{8} \right) = \log 8 \approx 2.079442. \end{aligned} \tag{17}$$

However, as assumed in the case of inter-firm networks, a firm is only active in the network if it has at least one link with another firm. As such, it can be deduced that the system is connected and will have at least 2 links. This excludes D_5, D_6, D_7, D_8 as possible choices for a prediction, making the probability of being correct $\frac{1}{4}$. The corresponding entropy is then

$$\begin{aligned} S &= \log 4 \\ &= 1.386294 \end{aligned} \tag{18}$$

Thus, the more uncertainty there is about the possible configuration of the network, the higher its entropy.

2.3 Configuring Networks

The fitness model from the introduction is based on a family of models called Configuration Models which mainly use information like degrees or weights to randomly generate networks. Configuration Models and their workings are discussed here. For a complete and detailed motivation and derivation of the weighted configuration see appendix B.

The idea of many of the models in this section is to estimate entries a_{ij} (or w_{ij}) via some probability p_{ij} derived from the maximising entropy framework, while using the knowledge of information like the degree and strength sequences which is encoded into the Hamiltonian i.e. exponent $H(G|\lambda)$ of the numerator of the expression in eq. (12). The parameters λ inside the Hamiltonian will then have to be estimated, which is done via a maximum likelihood method. Using the MLE on just the Hamiltonian allows estimating the system constraint applied only to the known portion of the network, while the rest of the system is still assumed to be as uniform as possible.

Using information on both the degree and strength sequences, one can reconstruct a network via the Directed Enhanced Configuration Model(DECM). For an $N \times N$ weighted network W and parameters $\alpha, \beta, \gamma,$ and λ the Hamiltonian takes the form

$$H(W|\alpha,\beta,\gamma,\delta) = \sum_{i=1}^{N} \left(\alpha_i k_i^{out} + \beta_i k_i^{in} + \gamma_i s_i^{out} + \delta_i s_i^{in}\right). \tag{19}$$

When plugging this Hamiltonian into the ERG formula, giving weight $w \in \mathbb{N}$ to the link between node i and j with probabilty q_{ij}, i.e.

$$P(W|\alpha,\beta,\gamma,\delta) = \prod_{i=1}^{N}\prod_{i\neq j} q_{ij}^{DECM}(w) \tag{20}$$

$$q_{ij}^{DECM}(w) = \begin{cases} 1 - p_{ij}^{DECM} & \text{if } w = 0 \\ p_{ij}^{DECM}(y_i^{out}y_j^{in})^{w-1}(1 - y_i^{out}y_j^{in}) & \text{else} \end{cases}$$

in which the shorthand is used:

$$\begin{aligned} e^{-\alpha_i} &\equiv x_i^{out} \\ e^{-\beta_j} &\equiv x_j^{in} \\ e^{-\gamma_i} &\equiv y_i^{out} \\ e^{-\delta_j} &\equiv y_j^{in} \end{aligned} \tag{21}$$

the resulting weighted link-incidence probability takes the form

$$\begin{aligned} p_{ij}^{DECM} &= \frac{e^{-\alpha_i}e^{-\beta_j}e^{-\gamma_i}e^{-\delta_j}}{1+e^{-\alpha_i}e^{-\beta_j}e^{-\gamma_i}e^{-\delta_j}-e^{-\gamma_i}e^{-\delta_j}} \\ &\equiv \frac{x_i^{out}x_j^{in}y_i^{out}y_j^{in}}{1+x_i^{out}x_j^{in}y_i^{out}y_j^{in}-y_i^{out}y_j^{in}}. \end{aligned} \tag{22}$$

further details can be found in Squartini et al. (2018).

To find suitable parameters for this model, the Maximum Likelihood Estimation method can be used on the Hamiltonian. This requires solving a system of $4N$ coupled equations which might prove computationally inefficient for larger networks. For Statistics Netherlands, the strength sequences are quantities known through sales and purchasing volumes. However, often due to privacy concerns, the degree sequences are not directly available. To address these issues a separate topological and weighted reconstruction method are used.

The DECM can be disentangled into seperate topological and weighted versions. For the topological part, the result is the simpler Directed Binary Configuration Model (DBCM). The DBCM uses only the in/out-degrees to reconstruct the link-incidence of (binary) network A with the Hamiltonian

$$H(A|\alpha,\beta) = \sum_{i=1}^{N}\left(\alpha_i k_i^{out} + \beta_i k_i^{in}\right) \tag{23}$$

which leads to (marginal) link-incidence probability

$$\begin{aligned} p_{ij}^{DBCM} = P(a_{ij}=1|\alpha_i,\beta_j) &= \frac{e^{-\alpha_i-\beta_j}}{c(\alpha_i,\beta_j)} \\ &= \frac{e^{-\alpha_i}e^{-\beta_j}}{1+e^{-\alpha_i}e^{-\beta_j}} \end{aligned} \tag{24}$$

To solve the problem of not knowing the node degrees, a 'Fitness Ansatz' is used: It is assumed that the topological attributes of a node are summed up by some node-intrinsic quality called 'fitness'. This relation has been observed in real-world trade networks, where a strong correlation between the constraints on the in/out-degrees w.r.t. certain non-topological in/out-fitnesses have been documented (Rachkov et al., 2021; Squartini et al., 2018; Cimini et al., 2015). This leads to the Fitness Model from the introduction, in the context of configuration models, it is called the fitness induced Directed Binary configuration Model(FiCM).

Let x_i^{out} and x_i^{in} quantify some intrinsic quality correlating to outgoing and ingoing connection preference. Specifically, say that there is a correlation between the degree sequences k and a certain predefined fitness x, then there exist linear functions f, g such that $e^{-\alpha_i} = f(x_i^{out})$ and $e^{-\beta_i} = g(x_i^{in})$, i.e. $e^{-\alpha_i} = \sqrt{a}x_i^{out}$ and $e^{-\beta_i} = \sqrt{b}x_i^{in}$ for certain a, b. Therefore, define a new parameter $z = \sqrt{ab}$. The FiCM then has link-incidence probability

$$p_{ij}^{FiCM} = \frac{z x_i^{out}x_j^{in}}{1+z x_i^{out}x_j^{in}}. \tag{25}$$

By another Maximum Likelihood argument it is possible to estimate the parameter z via

the number of links L in the network. By the constraint $L = \mathbb{E}L$ it follows that

$$\mathbb{E}L = \sum_{i=1}^{N} \sum_{j \neq i} \mathbb{E}a_{ij} = \sum_{i=1}^{N} \sum_{j \neq i} p_{ij}^{FiCM} \tag{26}$$

and thus z can be found by solving

$$L = \sum_{i=1}^{N} \sum_{j \neq i} \frac{z x_i^{out} x_j^{in}}{1 + z x_i^{out} x_j^{in}}. \tag{27}$$

However, the aggregate information on the number of links L in the network is also unavailable for Statistics Netherlands. Luckily, if one can sample on the number of links $\hat{L}$ in a subset of the network, say $G_s \subset G$, where $|G_s| = I$. The estimator z could then be found via

$$\hat{L} = \sum_{i \in I} \sum_{j \neq i \in I} \frac{z x_i^{out} x_j^{in}}{1 + z x_i^{out} x_j^{in}} \tag{28}$$

Note that finding this parameter only requires one equation to be solved, as opposed to the $2N$ equations of the DBCM. While being forced to use the FiCM as a result of limited information, in return a bit of computation time is won.

Using fitnesses for assigning links is not the only way of dealing with the unknown degree sequences, but unlike other configuration methods dealing with this lack of unavailable information, the FiCM incorporates a good-get-richer phenomenon. This will ensure that more attractive firms will be more likely to trade with eachother, resulting in clustering in the network. For instance, in Japan (Watanabe et al., 2013), because of less strict privacy rules the information on the degree sequences, associated with clustering, is more readily available. There it is found that the degree sequences are distributed according to a power-law. The FiCM incorporating a good-get-richer heuristic allows the configuration to reproduce the power-laws found in known networks.

It is often the case that strength sequences are a good approximation such that the sales and purchasing volumes could serve well as a general fitness (Squartini et al., 2018; Barrat et al., 2004).However, these fitnesses can incorporate other aspects as well. Statistics Netherlands has reasons to believe that lower relative distance between firms also adds to a higher connection probability (Dhyne and Duprez, 2016). Furthermore, SN has information on hard constraints that indicate wether certain industries trade with each other through input and output tables. A useful feature of the FiCM is that such additional ansätze can relatively easily be incorporated into the model.

Another component of the disassembled DECM, is calibrating the weighted network W. The Directed Weighted Configuration Model (DWCM) is a way to assign weights $w \in \mathbb{N}$ to each link according to a geometric distribution. If we take as the Hamiltonian in the ERG-model

$$H(W|\gamma, \delta) = \sum_{i=1}^{N} \left(\gamma_i s_i^{out} + \delta_i s_i^{in} \right). \tag{29}$$

then the weight of the link-incidence is assigned according to the probability distribution

$$Q(W|\gamma, \delta) = \prod_{i=1}^{N} \prod_{i \neq j} q_{ij}^{DWCM}(w) \tag{30}$$

where the marginal probabilities are

$$q_{ij}^{DWCM}(w) := (e^{-\gamma_i}e^{-\delta_j})^w(1-e^{-\gamma_i}e^{-\delta_j}). \tag{31}$$

There are a few ways this configuration method is expanded upon. First, the (weighted) configuration will be carried out conditional on the chosen binary configuration A, resulting in probability distribution $Q(W|\gamma,\delta,A)$. Secondly, the DWCM causes the distribution of the weights to depend solely on the strength sequences. In the spirit of the FiCM, it might be the case that other (empirical) properties also play a role in the weight distribution. As a national statistics institute, SN has access to a lot of auxiliary micro-data. Therefore, having the fitness feature in its toolkit is a considerable advantage. As such, another ansatz feature is imposed in the model. The Hamiltonian, carrying the information on our system will be defined by some target weights w_{ij}^* (carrying any ansätze) and corresponding Langrange Multipliers/parameters ζ:

$$H(W|\zeta) = \zeta_{ij}w_{ij}^* \tag{32}$$

Thirdly, we opt to use continuous weights $w \in \mathbb{R}_{\geq 0}$. Together, this will result in the distribution being exponential and defined by

$$q_{ij}(w|a_{ij}=1) = \zeta_{ij}e^{-\zeta_{ij}w} \tag{33}$$

The parameter ζ is determined via a generalized likehood method and given by

$$\zeta_{ij} = \frac{p_{ij}}{w_{ij}^*} \tag{34}$$

with p_{ij} the chosen binary marginal probability. This weighted configuration model is called the Conditional Reconstruction Method (CReM) (Parisi et al., 2020). As to choosing a target weight, many options are available. Exotic fitnesses are an option, but as with the binary fitness method, the strength sequences are still a reasonable starting point. As an example, take the Gravity Model as target weight distribution. This model is a deterministic method of assigning weights in a network as a proportion of their total strength. This method performs poorly when taking any topological features into account but generally performs well on the weighted configuration. Using the Gravity Model results in the following equation to be solved

$$\zeta_{ij} = \frac{W^\circ p_{ij}}{s_i^{out}s_j^{in}} \tag{35}$$

in which W° is the total strength of the system. Notice that this system requires the solution of $O(N^2)$ decoupled equations. In other words, by the nature of this configuration the parameters can be calculated in tandem with the assignment of links in the binary configuration method. This results in a relatively low computational load on the total reconstruction.

2.4 dynamic network reconstruction

The reconstruction of a network at repeated instances in time could in principle be done by independently applying the method to every time-slice, assuming complete independence. This is unrealistic because a trade relation being present at time 1 partly determines the probability that this connection exists at time 2. In other words some level of persistence is to be expected in the links.

One possible approach is to treat the constraints themselves as fundamentally constant for the two (or more) timeslices, but each is subject to some measurement error. This means that for all time slices the values of the aggregated quantities that serve as constraint are averaged, and those average constraints are then used to generate realisations. For every time slice a realisation is picked. This may produce some correlation or persistence of links between subsequent slices. Given the freedom that the constraints still allow for the reconstruction, the persistence of links, expecially for the lower strength nodes, will be very low.

Another approach is to determine a maximally correlated network for subsequent timeslices of the network. The notation for in- and out-degrees at two times t_1, t_2 is $k_i^{in}(t_1)$, $k_i^{out}(t_1)$, $k_i^{in}(t_2)$, $k_i^{out}(t_2)$, and as well the definitions are used:

$$\begin{aligned} k_i^{\mathsf{min},in} &= \mathsf{min}\left(k_i^{in}(t_1), k_i^{in}(t_2)\right) \\ k_i^{\mathsf{min},out} &= \mathsf{min}\left(k_i^{out}(t_1), k_i^{out}(t_2)\right) \end{aligned} \tag{36}$$

as the minimum of the in- and out- degrees for node i of time t_1 and t_2. Similarly the minimal strengths s are defined as:

$$\begin{aligned} s_i^{\mathsf{min},in} &= \mathsf{min}\left(s_i^{in}(t_1), s_i^{in}(t_2)\right) \\ s_i^{\mathsf{min},out} &= \mathsf{min}\left(s_i^{out}(t_1), s_i^{out}(t_2)\right) \end{aligned} \tag{37}$$

Now it is possible to reconstruct a network based on these minimal in- and out-degrees using the standard fitness approach. This step creates a base network to which further links can be added for either of the two times. The implicit assumption is that there is maximal persistence of links in the dynamic network. For instance if a node has degree 4 at time 1 and degree 2 at time 2 then the minimal degree is 2. This method makes sure that the persistent degree is also 2. In other words the 2 links present at time 2 are also present at time 1. The rest of the network then needs to be filled out for either of the two times. This can be achieved by:

$$\begin{aligned} k_i^{\mathsf{dif},in}(t_m) &:= k_i^{in}(t_m) - k_i^{\mathsf{min},in} \quad \text{for } m = 1,2 \\ k_i^{\mathsf{dif},out} &:= k_i^{out}(t_m) - k_i^{\mathsf{min},out} \quad \text{for } m = 1,2 \end{aligned} \tag{38}$$

as the differences, compared with the minimal values, of the in- and out- degrees for node i of time t_1 and t_2. Similarly the difference strengths s are determined:

$$\begin{aligned} s_i^{\mathsf{dif},in}(t_m) &:= s_i^{in}(t_m) - s_i^{\mathsf{min},in} \quad \text{for } m = 1,2 \\ s_i^{\mathsf{dif},out} &:= s_i^{out}(t_m) - s_i^{\mathsf{min},out} \quad \text{for } m = 1,2 \end{aligned} \tag{39}$$

These are then used in a similar way to augment the networks at t_1 and t_2 appropriately. In some sense this maximally persistent approach is the other extreme of the independent approach where any persistence of edges occurs purely by chance and therefore the persistence is as low as it can be. The actual realisation of the network will have to fall somewhere in-between. One can take that one step further by assuming for instance a characteristic time scale $\tau(s)$ for correlation decay: The strengths used to reconstruct the common sector of two time slices is then:

$$\begin{aligned} s_i^{\mathsf{min},in} &= \mathsf{min}\left(s_i^{in}(t_1), s_i^{in}(t_2)\right) e^{-|t_2-t_1|/\tau(s)} \\ s_i^{\mathsf{min},out} &= \mathsf{min}\left(s_i^{out}(t_1), s_i^{out}(t_2)\right) e^{-|t_2-t_1|/\tau(s)} \end{aligned} \tag{40}$$

Just as is done above the two time slices of the network are then augmented with additional edges in order to satisfy the actual network constraints for each slice. Application of these procedures is deferred to future work.

3 Methods and Measures

Lacking direct knowledge on the degree sequences but having access to large amounts of micro-data for firms by SN, makes any kind of fitness model a natural candidate for reconstruction. While many reconstruction methods mostly only require input on the degree and strength sequences, fitness models allow for the input of augmented sequences carrying information about (assumed) underlying structure in the true network.

To justify the choice of models, it is good to point out that an important reason for using the fitness model comes from a heuristic argument, which is empirically substantiated, combined with a restriction of the available information. For example, the true network is expected to be sparse and have degree sequences distributed according to a power-law. These are attributes that can be incorporated in the fitness models, without direct access to the degree sequences.

In general, if two models would heuristically both be good candidates, the Akaike Information Criterion (AIC) can be used to compare their performances. The AIC is defined by

$$AIC_m = 2d_m - 2\log \mathcal{L}_m \tag{41}$$

where m is the relevant model, d the number of parameters that need estimating for model m, and $\mathcal{L}_m$ the likelihood of the configuration probabilities. The idea behind the AIC is that adding extra parameters to the likelihood would increase its explanatory power w.r.t. the input data, but might lead to overfitting: explaining potential noise away by adding more parameters. The AIC is a trade-off between these two inputs, and the smaller the score the better the balance between numbers of parameters used, and explanatory power. The AIC represents a commonplace trade-off within statistical modelling where one wants to have relatively simple models, i.e. with few parameters, while also having good explanatory/predictive power.

It has been shown that for the World Trade Web (WTW) and the electronic Market for Interbank Deposits (e-MID) the DBCM should yield the best AIC score for the topological reconstruction of the network. The DBCM would also adhere to the heuristic requirements listed before, seemingly making it an ideal choice. However, the degree-sequences are unknown for the dutch business network. The use of this model, for the purposes of the interfirm trade network, is therefore infeasible(Squartini et al., 2018; Parisi et al., 2020).

Furthermore, SN divides the dutch firms into 650 commodity groups that can each consist of over 10.000 firms. This would result in a network described by a matrices with far more than 100 million entries. As such, the overall interfirm network of interest is presumably much larger than the bank or world trade networks. Therefore, there might be a considerable computational advantage in the use of the fitness model, as one only has to estimate a single parameter for the binary configuration instead of the coupled $2N$ equations the DBCM needs to solve.

Just as with the FiCM, SN's deterministic model also uses many assumptions based on correlations found in comparable networks. It derives the node degrees from a similar fitness, which is composed of the purchasing/sales volumes, certain industry scores, and a

measure of distance. Then it uses an empirical relationship between the in/out fitnesses and the in/out degrees of firms. The configuration is then carried forward by 'handing' out firm degrees in order of the firm strengths. A general overview of the method is found in appendix A and a detailed description can be found in Hooijmaaijers and Buiten (2019); Rachkov (2020); Buiten et al. (2021).

The method of deriving degrees is similar to the use of the Fitness Ansatz in the Fitness-model. Furthermore, the required input-data for the Fitness model is available to SN. A detailed comparison of the deterministic method versus the Fitness model has already been carried out in previous research by SN (Rachkov et al., 2021). Furthermore, the performance of first order measures like the True Positive/Negative Rate, and the accuracy and specificity are looked into. The main goal of this Paper is not to directly carry out a comparison with the deterministic method again, but to try and improve upon the performance of probabilistic reconstruction methods by themselves when applied to a network of trading firms. As one is generally blind to the real configuration of the system, in comparing the methods there is no true exact knowledge about what attributes are preferred over the other. It is, of course, still fruitful to do such comparisons. There is wisdom in knowing the differences of attributes produced by the different methods, as one can then choose the model that would be expected to better predict the truth. For the present case however, the reconstruction is mostly tested for its robustness and consistency, while also trying to improve its sampling scheme.

Algorithm 1: FiCM combined with CReM.

Data: s_i^{in}, s_i^{out}, I_{ij}, d_{ij} for all firms i, j in the desired commodity groups; Known links L_s on a sample of each commodity group $\alpha = 1, ..., M$

for $\alpha = 1, ..., M$ do
- N_α=# firms in α;
- $R = |\Omega_\alpha|$;
- $W_0 = \sum_{i=1}^{N} s_i^{in}$;
- for $r = 1, ..., R$ do
  - Solve $L_s = \sum_{i=1}^{N} \sum_{j \neq i} p_{ij}^{FiCM}$ to find z.;
  - Let A be an empty $N \times N$-matrix;
  - for $j = 1, ..., N$ do
    - for $i \neq j$ do
      - $b = \frac{p_{ij}^{FiCM} W_0}{s_j^{in} s_i^{out}}$;
      - draw q from an exp(b) distribution;
      - $a_{ij} \leftarrow q$ with probability p_{ij}^{FiCM};
  - $\Omega_{\alpha,r} = A$;

Result: Ensemble Ω_α of configurations for each commodity group $\alpha = 1, ..., M$

In earlier work, the fitness was chosen to be in line with the deterministic SN method, using the same distances d_{ij} and input/output tables I_{ij} in conjunction with the sales and purchasing volumes, which will also serve as in/out-strengths. As such the resulting fitnesses are given by

$$p_{ij}^{FiCM} = \frac{z s_i^{out} s_j^{in} I_{ij}/d_{ij}}{1 + z s_i^{out} s_j^{in} I_{ij}/d_{ij}} \tag{42}$$

The first addition, w.r.t. the previous work with the Fitness-model, is assigning weights, via the CReM method, to the network as well. The diagram 'Algorithm 1' above shows the algorithmic pseudo-code for these combined methods. To summarise, the idea of a fitness is to correlate some intrinsic node-specific quantity to its degrees. With this method, a good-get-richer phenomenon is incorporated in the produced networks. Fitter nodes would result in the corresponding firm to be more attractive for other firms to trade with. To tune the probability distribution that configures the network, the number of links in the network are estimated, according to a sample available to SN.

3.1 Network properties

One of the main advantages of using the FiCM is the emergence of certain higher order topological attributes being produced by the method. Certain patterns are seen in known networks like financial systems or other trade networks (Squartini et al., 2018; Cimini et al., 2015; Watanabe et al., 2013; Serra, 2020). One of these attributes is the earlier mentioned good-get-richer phenomenon, where the nodes with a higher fitness/strength are expected to be more attractive for other nodes to link with. One of the direct consequences of this, is that the degree distributions are expected to follow a certain power-law, such that there are a few nodes of very high degree and a lot with very low degree. More intricate, higher order[4)], attributes can also be investigated. Given the good-get-richer phenomenon, the average degree of the nearest neighbour (ANND) might also be relatively high. Since most nodes will connect to one of the high degree nodes, relatively fewer nodes will have a low ANND. Furthermore there are intricate webs of high density for a part of the network where the nodes of high degree are considered: the idea being that high degree nodes want to link themselves with other high degree nodes, producing a $clustering$ of nodes. In the spirit of validation described in section 3.3 it can not only be verified whether these patterns arise, but also check if these patterns arise consistently, such that the output mostly produces the same distribution for these attributes. This would mean that, if the model is well-specified, these emergent attributes are inherent to the system, not merely arising by pure chance.

In order to analyse the various patterns certain statistics are defined. As a reminder, in general A is the notation used for the adjacency matrix with entries a_{ij} and W is the corresponding weighted matrix, with entries w_{ij}. The degree distributions are then simply given by the degree sequences such that they correspond to the values given in (16). Furthermore, suppose the observed adjacency matrix is A^* and its corresponding weighted matrix W^* and links L^*. This results in

- The average nearest neighbour degree (ANND) of node i is the average degree of all i's directly connected nodes given by

$$k_i^{annd}(A^*) = \frac{\sum_{j\neq i} a_{ij}^* k_j^*}{k_i^*} = \frac{\sum_{j\neq i}\sum_{k\neq j} a_{ij}^* a_{jk}^*}{\sum_{j\neq i} a_{ij}^*}. \tag{43}$$

4) Higher order including properties not only conditional on the node itself, but also its neighbours or other objects in the graph.

- The average nearest neighbour strength (ANNS) is the ANND using the node strengths instead of degrees

$$k_i^{anns}(W^*) = \frac{\sum_{j\neq i} a_{ij}^* s_j^*}{k_i^*} = \frac{\sum_{j\neq i}\sum_{k\neq j} a_{ij}^* w_{jk}^*}{\sum_{j\neq i} a_{ij}^*}. \tag{44}$$

- The local clustering coefficient quantifies the fraction of 'triadic motifs', i.e. the number of triads (groups of 3 nodes that are connected) that are fully connected w.r.t. all triads. It is given by

$$c_i(A^*) = \frac{\sum_{j\neq i}\sum_{k\neq i.j} a_{ij}^* a_{jk}^* a_{ki}^*}{k_i^*(k_i^*-1)} = \frac{\sum_{j\neq i}\sum_{k\neq i.j} a_{ij}^* a_{jk}^* a_{ki}^*}{\sum_j \sum_{k\neq i.j} a_{ij}^* a_{ki}^*}. \tag{45}$$

- There is also a global variant of the clustering coefficient that calculates that number of closed triangles as a fraction of the total number of triplets. It is given by

$$C_i(A^*) = \frac{\sum_{i,j,,k} a_{ij}^* a_{jk}^* a_{ki}^*}{\sum_i k_i^*(k_i^*-1)}. \tag{46}$$

- The weighted (local) clustering coefficient is the the clustering coefficients weighted counterpart given by

$$c_i(W^*) = \frac{\sum_{j\neq i}\sum_{k\neq i.j} w_{ij}^* w_{jk}^* w_{ki}^*}{k_i^*(k_i^*-1)} = \frac{\sum_{j\neq i}\sum_{k\neq i.j} w_{ij}^* w_{jk}^* w_{ki}^*}{\sum_j \sum_{k\neq i.j} a_{ij}^* a_{ki}^*}. \tag{47}$$

  where the exponent is used to normalise the coëfficient.
- For a directed network the reciprocity is the fraction of nodes i, j that have both $a_{ij} = 1$ and $a_{ji} = 1$. It is given by

$$r(A^*) = \frac{\sum_i \sum_{j\neq i} a_{ij}^* a_{ji}^*}{L^*}. \tag{48}$$

One can specifically look at the directed versions of the ANNS and the ANND, by using s_i^{out}/s_i^{in} instead of s_i, and k_i^{out}/k_i^{in} instead of k_i. The ANND/ANNS and local clustering statistics give a value for each node in the system per realisation, whereas the global clustering and reciprocity are a single value per reconstructed network. It should be noted that the definition of the weighted versions of these higher order attributes is arbitrary. Consider the case when the weights are discrete , then the graph can be viewed as a 'multigraph' where each weighted edge between nodes i and j of weight $k \in \mathbb{N}$ can be viewed as k (binary) edges between nodes. The subsequent binary higher order attributes are then taken as the weighted versions. However, the implication here is that more importance is assigned towards higher volume edges. But who is to say that a weighted property is just a numerical multiplier of a topological property? The relative value of a weighted triadic motif w.r.t. a binary triadic motif should be dependent on the situation, therefore there are many possible versions of weighted higher order attributes. In the case of this paper a weighted version is chosen such that if all the weights would be equal to 1, their non-weighted counterparts would be retrieved.

Another statistic of interest is the cosine similarity measure φ_a. This is a first order weighted comparison measure. For a pair of weighted matrices W_1^a and W_2^a it is computed via

$$\varphi_a = \frac{I\,(W_1^a \circ W_2^a)\,I^T}{\|W_1^a\|_2 \|W_2^a\|_2}. \tag{49}$$

where I is the identity matrix, $\circ$ denotes entry-wise multiplication, and $\|.\|_2$ is the entry-wise L_2 norm. This can be seen as looking at the outcomes as a vector in an $N(N-1)$-dimensional plane and measure how much they point in the same direction by

looking at their inner-product. Therefore, it is a suitable way of comparing the performance of first order weighted properties of a network, without considering any topological features.

3.2 Sampling densities

One should be careful to derive estimates of the total network from a sample on part of the network. As mentioned earlier, exponential random graphs lack projectivity (Clauset et al., 2009). This means that the probability distribution on a sample will not be representative of the entire network. In the ERG setup the Hamiltonian H encodes various motifs into the network. Often such motifs incorporate a lot of conditional dependencies in the system. Any dyad i.e. any pair of vertices with or without an edge joining them, is not independent on the presence of other dyads in the system. To illustrate this, consider the good-get-richer phenomenon in the Fitness-model. The assumption is that the probability of a link-incidence is higher for nodes that already posses several connections to other nodes. Then these already existing links have a probability dependent on other nodes as well. The resulting dependency structure gets very complicated for bigger graphs. The consequence of this dependency problem is that the Hamiltonian is not a sufficient statistic for the probability distribution on the entire network.

This should make the sampling scheme, employed to tune the fitness model, suspect to potential bias. Aside from theoretical suspicions, SN has reason to believe the data they possess on the density contains a bias. The data on the business activities on part of the network originates from the commercial data provider Dun & Bradstreet. The concern is that the sample at hand overestimates the actual density of a network, mainly because bigger firms are more likely to be present in these kind of surveys.

Algorithm 2: Resampling scheme

Data: A reconstructed matrix W^*, desired number of resamples B, size of subset of firms n.
Let D be a zero-vector of size B;
for $i = 1, ..., B$ do
  randomly sample an $n \times n$ block W_n from the matrix W^*;
  $d = \frac{\text{Links in } W_n}{n(n-1)}$;
  $D_i = d$
Result: A vector D of resampled densities

In Musmeci et al. (2013) an empirical study is performed on deriving topological properties from an ERG using sampling. Resampling on a network, and taking averages leads to an accurate estimation of the density of a network. Of course, samples on density are not easy to obtain[5)], thus one can not hope to endlessly acquire (re)samples on a network. When confronted with a lack of samples often bootstrap techniques are employed where one takes resamples from the available data. While this method can be computationally demanding, if the available sample is representative of the population or

5) As exhibited by SN's use of a commercial dataset.

system, bootstrap attains accurate statistical results. In the case of the FiCM's density (re)sampling method, the problem of bootstrapping is in its suspected bias.

Nevertheless, there is some investigation possible on the sampling scheme, where adjustments could be made to detect or alleviate biases present in the sample available. A test on the sampling scheme can be simulated by constructing a single network using the FiCM and SN-data. Then, the sampling scheme can be performed on the network and, together with the SN-data, an ensemble can be produced from which the densities can be analysed. The idea is that resamples or 'bootstrap samples' may simulate a survey taken amongst firms about their possible activities within the business network. The size of the sample is then denoted as n. The scheme looks as shown in the pseudocode above (Algorithm 2).

Note that these tests are very empirical in nature. Indeed, there is as of yet little asymptotic theory available when it comes to ERGs. An overview of the state-of-play is (Kolaczyk, 2007), and in the decade since that book has come out there has been some incremental progress in this regard, but nothing sufficiently substantial yet. As such, this resampling scheme is mostly a tool of inquiry for surprising/unsurprising results from the model and to further test some consistency features (albeit empirically). Little hard mathematical theory can be used to show results as the therefore necessary theoretical quantities are unknown.

3.3 Testing against the unknown

After running the reconstruction procedure, its performance is evaluated. Whilst many statistical indicators are available, there is little topological information about the real configuration for testing. This is a frequently encountered problem of reconstructing networks, which is of course the cause for wanting to use reconstruction methods in the first place. Earlier applications of the FiCM/CReM have been on the World Trade Web or Financial banking services, where more information is available. Also in some other EU countries, transactions data are registered for tax purposes so that methods could be better tested in future. In these situations it is possible to construct the matrix P containing all link/weight-probabilities p_{ij}. The usual method of testing the reconstruction is to compare the reconstructed networks against their theoretical quantities derived from the probabilities p_{ij}. This way one would learn about the robustness, consistency and accuracy of the method. In the present case, this method is computationally demanding since the method has to deal with commodity groups of more than 10.000 nodes, and the only reason it is possible to create small ensembles with relatively small memory is because the matrices are sparse. This sparsity is not of help in the computation of P, or the derivation of its attributes, as each probability needs to be calculated and assigned to an entry in the matrix, making it slow while also requiring a lot of memory. To still be able to make any inferences concerning the consistency, some ideas from non-parametric statistics and machine-learning can be borrowed.

In machine-learning the notions of recognition, fitting, and prediction of patterns in data are central. There, the goal is often to 'predict' or identify specific sought after patterns. The models are trained on data, and (part of) the art of machine-learning is to teach a model how to recognise patterns whilst simultaneously not over-fitting it on the available data. In order to test against new information one employs a method called validation.

Here, the available data is split into a training and a validation/holdout set. The idea being that the model is trained on the training data, and then tested on the validation data (Shalev-Shwartz and Ben-David, 2014).

This validation idea is adopted, whilst being mindful of the differences between machine-learning and the Exponential Random Graph context. Firstly the data should contain patterns. One could think of for instance the fraction of closed triangles compared to the maximum number of possible closed triangles given the number of links, or the fraction of stars, or other such substructures, but exactly which patterns should be observed is unkown. Secondly, similarly to many machine-learning cases, there are a lot of data available [6], but the heterogeneity of the systems of interest makes any kind of over-fit (or bias) unwanted. The very point of using entropy maximisation is to not exacerbate any 'prediction' bias present in the input-data.

As such, when carrying out a validation on the ensemble of configurations presented as matrices $\Omega_R \in \Omega$[7] produced by the chosen model. Ω_R will be split up into training set $\Omega_{R,1}$ and validation set $\Omega_{R,2}$. If the model is well-specified, Ω_R consists of multiple possible truths, while hopefully also containing the actual real-world true configuration, albeit within some bandwidth. Now, one can take any contextually useful quantitative statistic T and regard $T(\Omega_{R,1})$ as the hypothetical true quantity of this statistic. Then it can be tested how well the set $\Omega_{R,2}$ performs w.r.t. the hypothetical truth $\Omega_{R,1}$ i .e. how close $T(\Omega_{R,2})$ and $T(\Omega_{R,1})$ are.

However, at this point it is important to be clear about what exactly is gained in knowledge. If $\Omega_{R,2}$ is performing well in this scenario, that only means that it performs similarly to $\Omega_{R,1}$. So if $\Omega_{R,1}$ is a wrong output, the only lesson learned is that $\Omega_{R,2}$ is equally wrong. However, here equality is actually useful information. When the split up-sets are similar in behaviour it means that the configuration method is robust, i.e. giving us consistent outcomes. This is why it is important for the model to be well-specified. If the model contains the true distribution/configuration of reality and if the model produces robust, low variance and consistent outcomes, then the model will have a high probability of producing an ensemble containing something close to the real network configuration.

4 Results

Here some highlights of the research will be showcased. A detailed description and investigation of the results can be found in Kayzel (2022). The results are acquired from producing ensembles of reconstructions from 5 different commodity groups and analysing them. Due to computational constraints the number of groups to be investigated is limited, but the groups chosen are somewhat diverse such that it can be assessed how well the reconstruction performs on a wide variety of commodity groups. Furthermore, the size of the chosen commodity groups is reduced by leaving out any firm with a sales or purchasing volume smaller than €10.000.

6) In case of the strengths, one could say SN possesses the knowledge of the entire population.
7) Here Ω denotes the relevant sample space of the probability space enveloping the reconstruction procedure.

The commodity groups are considered as closed systems, such that the in/out-volumes are the same. If $s^{out} \neq s^{in}$, then the direction with the bigger total volume is reduced by a fraction $f = \min\{s^{in}, s^{out}\}/\max\{s^{in}, s^{out}\}$, making their total strengths equal. Since the method requires assigning likelihoods proportional to the strengths it is required that there is a normalisation so that the sum of all likelihoods is 1. This holds separately for in-strengths and out-strengths, which in practice implies a need for multiplying by this factor so that the requirement can be met.

There are several reasons why such a difference can arise between input and output sales volume in the data at all. First of all the cutoff for low sales volumes can generate such a discrepancy if for instance there are many more firms with small sales volumes than there are with small purchase volumes, or vice versa. Such asymmetries will balance out over the full dataset, but not a dataset where a cutoff is applied. Secondly, trade with firms outside of the country is normally not explicitly accounted for, which can cause similar asymmetries as well. Thirdly, if a transaction is near the beginning or end of any given accounting period (a month, a quarter, a year, or any other time frame) either the sale or the purchase of a commodity might have been or will become registered in the adjoining time frame so that there is an apparent discrepancy in sales and purchasing volumes.

For each commodity group $a = 1, ..., 5$, the reconstruction procedure of the FiCM and the CReM is carried out $R = 20$ times, giving an ensemble Ω_R^a, where $|\Omega_R^a| = R = 20$ for all a. Most of the time the ensemble is splitted and validated, i.e. computing averages of the statistic of interest and compare the outcomes for each half of the ensemble. In some cases, the topological and weighted performance are separately evaluated, but in cases like ANNS or the weighted clustering coefficient, the observed quantity says something about both.

4.1 First and second order attributes

The weighted reconstruction by the CReM can be evaluated by plotting the produced strength sequences against the known strength sequences. Below this is shown for the commodity group Barley.

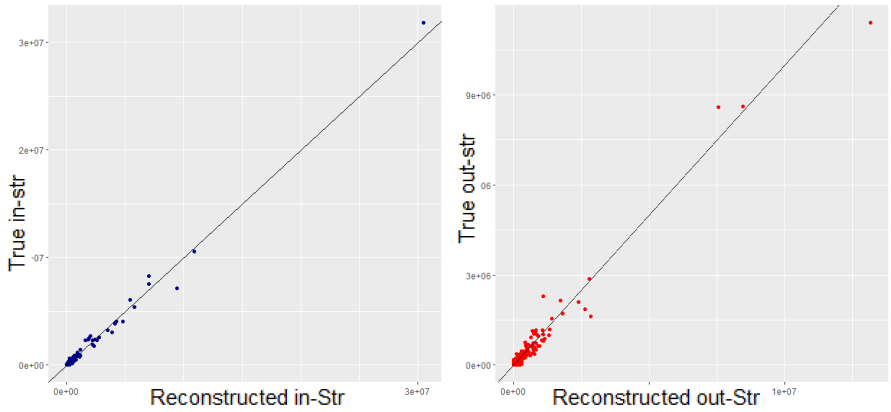


Figure 4.1 Commodity group 1 (Barley): reconstructed strengths per node by CReM model of a single realisation vs the known true values, with the 45° line plotted for reference.

To quantify this similarity the Pearson correlation coefficient is used. This is a good measure of correlation when the relationship is assumed to be linear. In most cases the

Pearson correlation coefficients for the strengths versus the reconstructed strengths are very close to 1. The exceptions to this rule are the in-strengths of the commodity groups Water and Steel. In the case of steel, this could be caused by the presence of a high strength outlier in the group, causing discrepancies. In the case of water it might be due to the relatively small number of suppliers in this group.

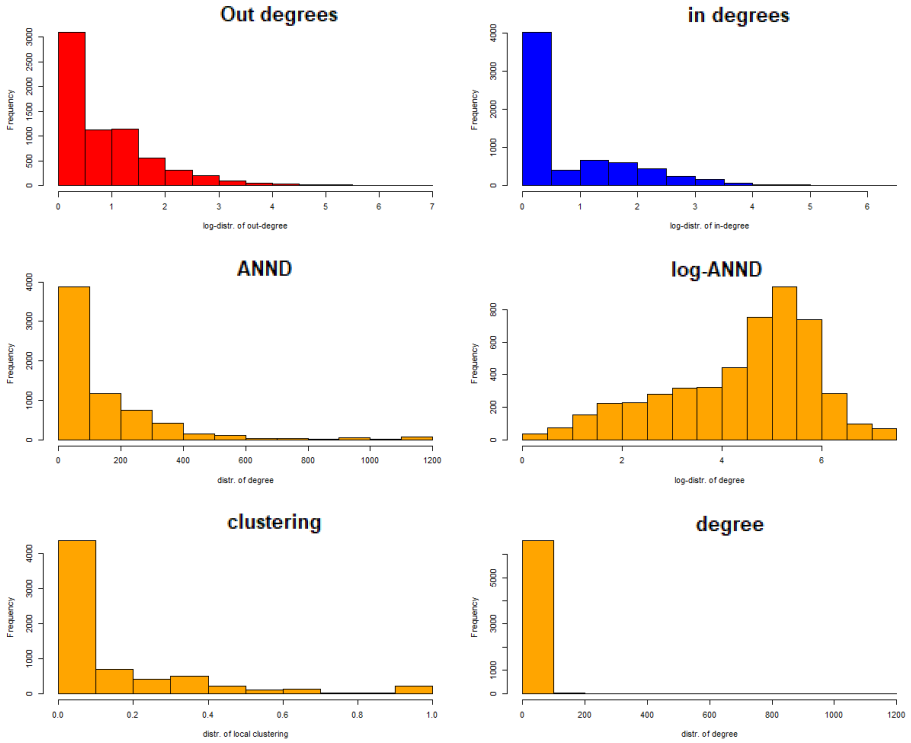


Figure 4.2 Distributions of a realisation of the commodity group 3 (books), blue and red are in/out direction resp. whereas yellow indicates no direction.

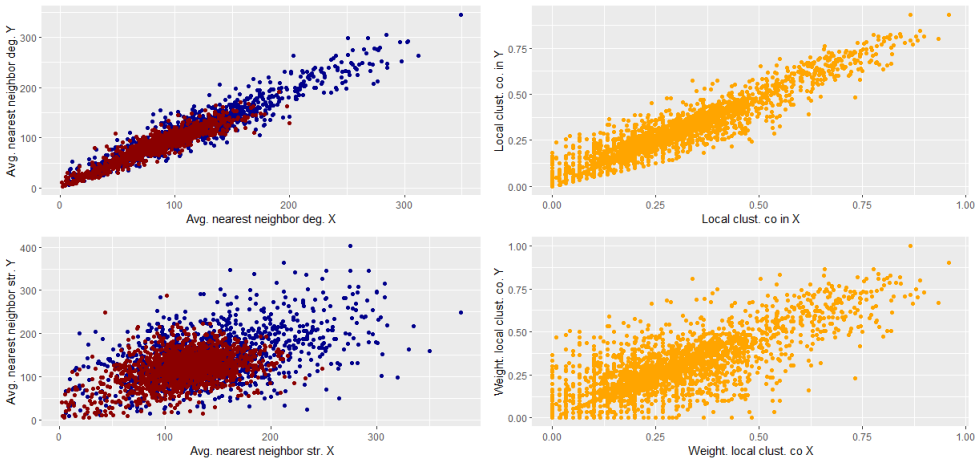


Figure 4.3 Second order attributes of the commodity group videogames, results for the splitted samples X and Y plotted against each other. Top left: the average nearest neighbour degree, bottom left, the average nearest neighbour strength, top right: the average clustering coefficient, bottom right: the weighted average clustering coefficient. Red dots indicate an out-direction, blue is in, orange is undirected (for clustering).

Recall that from earlier research it is known that the power-law degree distributions are observed when using the FiCM as the binary reconstruction method. This distribution has been observed in real-work interfirm networks and it is attempted to recreate this

observation through the use of the Fitness Ansatz. So as expected, it can be observed in the reconstruction of the commodity groups investigated here, as seen in figure 4.1. While these histograms only show the distributions of a single realisation, these patterns can be observed across all commodity groups investigated.

Perhaps more unexpected is the behaviour of the higher order properties. If these are also consistently reproduced in the reconstruction method then these reconstructed distributions could be compared to those observed in real networks to see if they match.Remember that the theoretical values for the higher order properties are unavailable, since calculating the corresponding attributes of expected values into a matrix P^{FiCM} consisting of p_{ij}^{FiCM} is too costly due to memory allocation limitations. This is where the splitting and validation scheme described in section 3.3 comes into play. If the higher order quantities of two splitted ensemble halves are high correlated, there is justification for the hypothesis that they consistently reproduce these properties. In Figure 4.3 these relationships are plotted out for the commodity group of video-games.

It is expected that the weighted properties would display more volatility. As mentioned, the weighted attributes are also carrying topological information, thus the resulting weighted higher order properties will always be less consistent than their mere topological counterparts. What is remarkable is the relatively consistent reconstruction of these higher order properties in the networks. Note that the sampling scheme carries with it very little information about the systems topology, but the fitnesses used to recreate the good-get-richer phenomenon does translate higher order properties into this configuration.

4.2 Sampling

To evaluate the sampling methods used in the FiCM a single realisation W_0^a from the reconstruction procedure will serve as a synthetic known network to perform the statistical procedures on and compare them to the synthetic truth. The statistic of interest is the density. In the chosen method a sampled density is used for the tuning of the parameter z used in the reconstruction. When simulating this for the synthetic network, does this procedure induce a lot of bias, or is the influence on the resulting construction manageable?

As a synthetic network the first realisation of the ensemble from the commodity group videogames is taken, consisting of 3237 firms. This commodity group is chosen in particular, because it is one of the more consistently performing ensembles. Also, it is relatively small, so easier to simulate on. For the purpose of the density, the focus on one realisation is sufficient since the constraints on the links enforce that these values will be reconstructed with very little variance. For example, in the case of the videogames ensemble

- With 17216 links the standard deviation of the links in the ensemble is 117, which indicates relatively minor variation on a network of 10474932 possible links. The resulting density d^* is then given by

$$d^* = \frac{L^*}{N(N-1)} = 0.001643543.$$

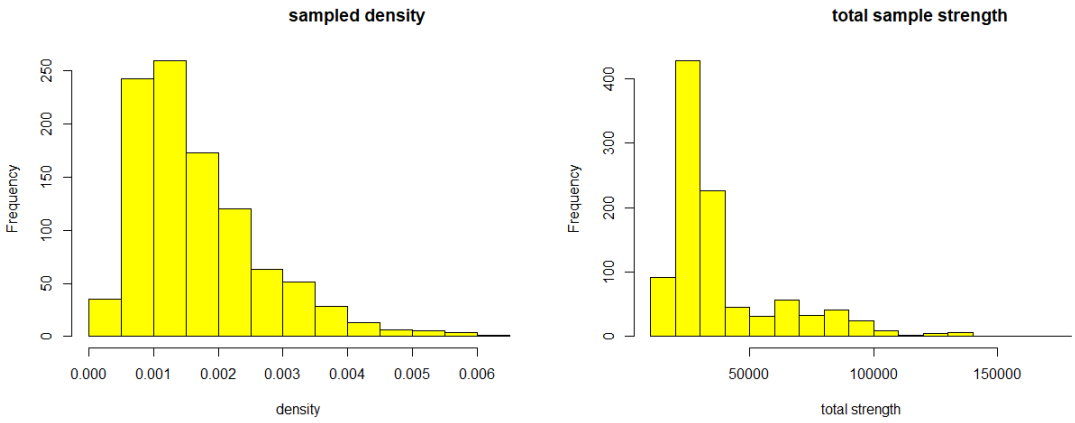


Figure 4.4 Left: the histogram of the densities obtained from resampling 1000 times on a synthetic network W^* (drawn from the commodity group Videogames). Right: histogram of the total true strengths of the samples.

- The sample used for the reconstruction is of size (i.e. the number of possible links) 26866, in which $L^*_{samp} = 131$ links where found giving us an implied density of $d^*_{samp} = 0.00487605151$.

What stands out is that the implied density is almost 3 times higher than in the resulting reconstruction. So does the model receive little influence from this sample, or does a lower implied density result in an even lower realised density? To test this, view the reconstruction W^* as a true (synthetic) network and employ the resampling procedure, described in section 3.2. A subsample of the firms of size $n = \sqrt{26866} \approx 163$ is taken and resampled $B = 1000$ times. This results in a list D of 1000 sampled densities where the min $D = 0.000265$, min $D = 0.006286$ with mean $\bar{D} = 0.00168$. Performing the FiCM reconstruction with parameter z tuned by this lower density resulsts in a network with 1405 links, whereas z tuned by the higher density produces a network with 22080 links. So the density of the sample does matter quite a bit in the resulting configurations.

It is then problematic that, as seen in figure 4.5, the sampled densities vary a lot. However, taking the mean of D results in a density fairly close to the true density of the network. This is in agreement with both the constrained reconstruction method and the random sampling scheme proposed in section 3.2 and Blagus et al. (2017). This result also accurately approximates the true density when the number of resamples is drastically reduced to $B = 50$, giving a mean density of 0.0016784. When lowering the desired number of samples and considering confidence intervals a studentised bootstrap is used to verify the accuracy of the resampling schemes means (Asmussen and Glynn, 2007). For instance, when only resampling 10 times a mean of 0.001348 is achieved, but when considering its confidence intervals the resulting mean would be quite inaccurate. However, even these inaccurate confidence intervals already give a closer approximation of the true density of the synthetic network than the SN sample, obtained from Dun & Bradstreet, does. Detailed results of various density estimations and resamples can be found in table A.

Some notes on Table C:

- For large sized resampling, reliable asymptotic confidence-intervals can be computed

Table C	n=5	n=20	n=50	n=163
B=10	no links	0.00289	0.00139	0.00128
CI-st.bootstrap	sampled	[-.00334, .00913]	[-.00203, .00399]	[.00002, .00248]
B = 20	no links	0.0025	0.00141	0.00176
CI-st.bootstrap	sampled	[-.00886, .0103]	[-.00344 .00486]	[.00016, .00322]
B= 50	0.004	0.00189	0.00152	0.00181
CI-st.bootstrap	[-.047,0.0033]	[-.00527, 0.00756]	[-.00151, .00435]	[.00003, .00349]
B=1000	0.00155	0.00154	0.00168	0.00165
Asymp. CI	[.00085,.00224]	[.00136, .00172]	[.00156, .00179]	[.0159, .0171]

as those sets contain enough values for good approximation. However, in reality acquiring this many samples is unfeasible.

- Note that this table also illustrates a pitfall of using a studentised bootstrap. Sometimes the lower bounds of the density goes below zero, which is nonsensical. This is due to the methods assumption that these negative outliers have not yet been realised while in fact these outliers are simply impossible events. This could also happen with asymptotic intervals, but in this case, when the resamples B are large enough for the asymptotic CI to be effective the bandwidths are narrow enough for it not to be an issue.
- Especially when using a low number of resamples on a smaller sampling size, these confidence intervals should be taken with a large grain of salt. But notable improvements are made by merely increasing the sample-sizes.
- Acquiring samples for commodity groups is no small feat, but with the unavailability of the node-degrees, a random sampling scheme might be a solution to alleviating biases present when using only one sample. A trade-off could be made for using smaller sample-sizes but applying the random selection scheme from Blagus et al. (2017). Note that the sampled density of the actual subset used in the reconstruction overestimates the density of the system by quite a bit, even relative to some alternative sampling schemes proposed here.

5 Discussion

The overall goal of the methods described in this paper is to reconstruct an interfirm trade-network, using partial information. Previous methods employed by SN to achieve this involve assumptions backed up by empirical studies. Using these assumptions to infer on the network, may lead to unwanted biases, which was shown in a previous study on this topic (Rachkov et al., 2021). Furthermore, these methods are deterministic, leaving room for only one inferred reconstruction. Since the reconstructions are an example of mass-imputation, in this case of the existence of links, all the problems associated with such efforts have been the subject of intense debate within SN in the past. Relying on only one method, and only one realisation of a reconstruction is particularly undesirable, so even if the deterministic reconstructions might appear plausible, it is very important to augment those results with alternative valid reconstructions.

Probabilistic methods exist that produce entire ensembles of possible configurations, while alleviating biases. Configuration models, employing maximum entropy, configure

the network using the available information but remaining maximally random about the unknown part of the network. These methods have been applied and tested on banking and international trading networks (Squartini et al., 2018; Cimini et al., 2015). In limited ways it has also been tested on interfirm trade-networks. For the specific purpose of reconstructing the network of a commodity group, the FiCM and CReM can combine the maximum entropy ideas with the assumption made in the deterministic method. The assumptions are used to base inferences on the known parts of the network, while trying to maintain maximally (uniformly) random on the rest. A sample of the network is used, to tune the parameter in the probabilistic configuration.

It is the aim of this discussion paper to build stronger foundations on these maximum entropy methods. In configuring networks, the topological reconstruction is a harder problem than the weighted reconstruction, partly because there is less information about the topology of the network. For the weighted reconstruction, often a degree corrected gravity model or Iterative Proportional Fitting is employed. This is a deterministic way to distribute weights conditional on some binary configuration of the network, and they have generally good performance (Squartini et al., 2018). Here it is proposed to use the probabilistic method CReM, that distributes weights conditional on some binary configuration using a exponentially distributed link weight. While reconstructing the weighted allocation of a network is not the hard part of the problem of producing business production networks, it is good to have an option to allocate weights that can naturally work with auxiliary data input. Using these probabilistic methods, ensembles of possible configurations are produced and evaluated, while limiting computational requirements.

5.1 Networks as production statistics

It is important to note that the employed models accurately reproduce the known information in the network, but often for the unknown part there is little knowledge on the desired output. Studies have been performed on the trading networks in Belgium and Japan that see a power-law emerging in the degree distributions of these networks, and these power-laws can also be seen in our outcomes. However, for many higher order properties no empirically validated knowledge on their distributions is available. Ultimately, the Fitness Ansatz used to infer on the structure of the system tries to encode a preferential attachment process called the good-get-richer phenomenon that causes this power-law of the degree distribution to appear. The somewhat indirect consequence of this is that some higher order properties seem to be consistently reproduced as well. The fact that these structures appear indicates there are emergent properties present in the system. It is currently unknown whether the structures that emerge from this model coincide with the structures observed in the real world trading networks of Belgium and Japan (or other known interfirm trade-networks). The question remains then if these observed higher order structures are a true indirect emergent property of the model, and if they are suitable for the prediction of the network dynamics.

With the power-laws resulting from the preferential attachment, it is expected that a few firms attract a relatively large part of the connections in the system. As such, it can be seen that the network has some high activity in the tails of its distribution. When looking at higher order properties the activity surrounding nodes is analysed, instead of the nodes themselves. Presumably, there will be a few neighbourhoods or clusters of nodes that also

have a relatively high activity w.r.t. the rest of the system. As such one could hypothesise that these properties will also contain high activity in their tails. Even more so, since the activity registered involves a whole cluster of nodes, the tails will be fatter than with the first order attributes. These fatter tails can be observed in the case for the commodity-group networks reconstructed in chapter 4. In truth, it is necessary to empirically validate such hypotheses, in order to draw any substantial conclusions, but it is reassuring to see the model behaves as expected.

What can be investigated, is whether the model produces these second order results consistently. In general, especially the binary configuration seems to be fairly consistently produced. The varying results are presumed to be caused by the commodity-specific structure encountered in each. For example the commodity group water is unique because it has very few suppliers, and low link-reciprocity which might cause inconsistency in the clustering. Moreover, the commodity-group of steel performs relatively mediocre at consistently producing most attributes. This commodity group contains one very strong firm, that seems to be causing issues. While no hard conclusions can be drawn, it is promising to see that the FiCM+CReM seems to produce some notion on the structure of higher order attributes. This is an advantage over a deterministic method, which lack the emergence of these structures.(Cimini et al., 2015). A detailed study of higher order behaviour of the network constructed using the rule-based method of SN is yet to be carried out.

Lastly, with the simulations run on the sampling scheme, some alternatives are proposed to the current sampling method employed. Statistics Netherlands is aware that the available sample on the number of links in the system, from Dun & Bradstreet, is biased. The firms that are included more regularly in samples to be surveyed are the firms that in macroeconomic terms matter more to the Dutch economy, i.e. larger firms with a large turnover or with many employees. The available sample similarly contains predominantly large firms. One might therefore hypothesise that such samples show stronger clustering than the population as a whole. Even so, merely due to the lack of projectivity inherent in the ERG-framework, any single random sample would be biased with high probability. Ultimately it would be ideal if the node-degrees were known. However, this would require highly detailed information on firms' activity, which is an administrative burden to collect unless it is also required for, for instance, taxation purposes as is the case in Belgium for instance. While samples are hard to collect, table C shows that it might be feasible to perform relatively small surveys with $400 - 2500$ firms in a system, and do this multiple times for a randomly chosen group of nodes. For example, when conducting surveys of size 2500 and doing it 10 times, in the end the number of firms surveyed is 25000, which is still less than the sample used in the commodity group of Video-games, which used 26886 firm-relations as a sample. With the random sampling scheme the average density would be a closer approximation of the true density of the network than the single large sample would imply. For SN, it will be interesting to consider such options. However, they are limited in acquiring these samples. The samples are acquired from an external company, Dun & Bradstreet, and it might not even be feasible to conduct that many surveys due to cost limitations. Furthermore, it is easier to survey a few large firms with many links, than to survey a lot of small firms with few links. In spite of the probable lack of alternatives, it is still important to investigate the biases present in the current available sample.

5.2 Limitations and Future research

An important and as of yet unsolved problem is the current closed nature of the interfirm trade network. The configuration models work on the assumption that they are closed systems, meaning that the ingoing and outgoing links' total weight should be the same. The models uses this fact often when deriving the link-incidence probabilities or maximum likelihood constraints. In reality the strength sequences derived from the supply/use and input/output tables by SN often lack this property.

It should also be stated that the use of a Fitness Ansatz in combination with link-sampling is done by necessity, instead of effectiveness[8]. In the end, the FiCM tries to approximate the degrees in the system, and then uses the marginal probabilities of the DECM/DBCM to assign links. As such, it would result in a better performance if these degrees where exactly known. As stated earlier, it is unrealistic to assume that this information would be available. It is shown in Squartini et al. (2018); Parisi et al. (2020); Cimini et al. (2015) that if one were to sample the in/out-degrees of a subset of firms (instead of the links) it would also be sufficient to reconstruct the networks. In the case of banking and international trade-networks it seems that this method outperforms the link-sampling method. However, the banking and commodity trade networks are very different so it would be wise to not draw conclusions on these performance differences too hastily.

Even so, the current sampling method could be improved by slightly adapting the way the sample is acquired. If such practical solutions prove to be unavailable, one could also try and alleviate the sampling bias by inferring some structure on the distribution of links on the network as proposed in Squartini et al. (2017). The sample could then be readjusted according to this structure. While only the available strength of a sample might not be sufficient to arrive at such a structure, further research can be done to for the pursuit of such a structure. Perhaps the fitnesses themselves can prove to be better predictors of the link-structure.

As for the weighted reconstruction, there are other weighted configuration methods to be investigated. The CReM seems to work well with the Fitness model on a (admittedly) heuristic level, but other models might prove to give better results. Even when sticking to the CReM model, one could investigate if other target distributions than the gravity model give different outcomes, and which role auxiliary data could play in the improvement of this method.

The research done on commodity-groups so far only focuses on each individual group. It would be of great interest to SN to also look at the network from a multiplex perspective where multiple commodity-groups are considered as part of high-layer of industries or countries etc. For this, it is of import to consistently predict that a link between two firms in a commodity-group remains the same when seen from a different layer.

For all its supposed merits there are some important shortcomings unique to the probabilistic method. First, they do not always produce connected configurations. Each node in a system should have at least one link, otherwise their activity in the system is nonsensical. This problem can be alleviated by considering an ensemble. Each individual

8) Although it does provide some computational advantage.

realisation may contain links that are disconnected, yet over the entire ensemble there is always a realisation to be found where that firm does contain a link. However, this is where the deterministic method of SN has an advantage, as it distributes links in a way that ensures each node gets assigned at least one link. Secondly, due to the probabilistic nature of the method it currently requires more computation time than the deterministic method. This is partly due to the fact that the method has to produce multiple networks to compose an ensemble, which is needed for any statistical inference. Lastly, SN argues that the probabilistic nature of the FiCM and CReM make the method inconsistent and therefore unsuitable for the use of production statistics (Buiten et al., 2021). While there is some truth in this, it is important to point out that the constraints used in these methods make the outcomes probabilistically consistent. That is to say that the ensemble averages converge in probability to their theoretical expectations. Furthermore, there are inaccuracies in both probabilistic and deterministic methods that make any point estimate very unlikely to represent the real world scenario. It could then prove advantageous to not speak about a single point estimate, but to speak in terms of bandwidths and confidence intervals, that could contain the truth with high probability.

As a final note it should be stated that the reasons for pursuing a reconstruction of configuration of these networks is only briefly alluded to in the introduction. The entire analysis of dynamics in time, risks, centrality or vulnerability (to name a few) is not discussed extensively here, although some initial steps in this direction are presented in section 2.4. This is the subject of further research that is still in progress. In order to perform these analyses on a network, it is of great importance that the networks considered are grounded in some truth. Therefore, the reconstruction procedure should be considered as an important first step when trying to infer information about a system's structure. It is then interesting to consider that the ensembles produced by these methods contain possibilities of the configuration, and because of this it is possible to ascertain the risk of many possible scenario's instead of being restricted to only one determined world.

References

Arthur, W. (2021). Foundations of complexity economics. Nature Reviews Physics 3, 136–145.

Asmussen, S. and P. Glynn (2007). Stochastic Simulation: Algorithms and Analysis. Springer.

Barrat, A., M. Barthélemy, R. Pastor-Satorras, and A. Vespignani (2004). The architecture of complex weighted networks. PNAS 101(11), 3747–3752.

Blagus, N., L. Šubelj, and M. Bajec (2017). Empirical comparison of network sampling: How to choose the most appropriate method? Physica A: Statistical Mechanics and its Applications 477, 136–148.

Buiten, G., E. de Jonge, G. Mooijen, S. Hooijmaaijers, and P. Bogaart (2021). Reconstruction method for the dutch interfirm network including a breakdown by commodity for 2018 and 2019. Technical report, Statistics Netherlands (CBS).

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

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009, nov). Power-law distributions in empirical data. SIAM Review 51(4), 661–703.

Dhyne, E. and C. Duprez (2016). Three regions, three economies. Technical report, National Bank of Belgium.

Gandy, A. and L. A. Veraart (2016, oct). A bayesian methodology for systemic risk assessment in financial networks. Management Science 63(12), 3999–4446.

Gandy, A. and L. A. Veraart (2019, jul). Adjustable network reconstruction with applications to cds exposures. Journal of Multivariate Analysis 172, 193–209.

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

Kayzel, J. (2022). Network reconstruction: Producing ensembles of possibilities. Technical report, University of Amsterdam.

Kelly, F. and E. Yudovina (2014). Stochastic Networks. Cambridge University Press.

Kolaczyk, E. (2007). Statistical Analysis of Network Data. Springer.

Kullback, S. and R. Leibler (1951). On information and sufficiency. Ann Math Stat 22, 79–86.

Musmeci, N., S. Battiston, G. Caldarelli, and M. Puliga (2013). Bootstrapping topological properties and systemic risk of complex networks using the fitness model. Journal of Statistical Physics 151, 720–734.

Parisi, F., T. Squartini, and D. Garlaschelli (2020). A faster horse on a safer trail: generalized inference for the efficient reconstruction of weighted networks. New Journal of Physics 22.

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

Rachkov, A., F. Pijpers, and D. Garlaschelli (2021). Potential biases in network reconstruction methods not maximizing entropy. Technical report, Statistics Netherlands (CBS).

Serra, P. (2020). Lecture notes non-parametric statistics. Technical report, Tech.Univ. Eindhoven.

Shalev-Shwartz, S. and S. Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

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, 1–47.

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

Squartini, T. and D. Garlaschelli (2017). Maximum-Entropy Networks Pattern Detection, Network Reconstruction and Graph Combinatorics. Springer.

Watanabe, H., H. Takayasu, and M. Takayasu (2013, feb). Relations between allometric scalings and fluctuations in complex systems: The case of japanese firms. Physica A: Statistical Mechanics and its Applications 392(4), 741–756.

Appendix

A The deterministic SN-method

To setup input data of firms with their trade volumes there are numerous steps involved. The firms are selected from the Statistical Business Register, and supply and use volumes are classified per commodity to arrive at a firm to firm input and output per commodity. A detailed review of the reconstruction method employed by SN can be found in Rachkov et al. (2021); Hooijmaaijers and Buiten (2019); Buiten et al. (2021), but an overview of the way the binary configuration is acquired is given here. Some of the scores defined here are also utilised in the fitnesses of the FiCM[9].

For the reconstruction method used by SN, consider a single commodity group α. The firm-level marginal strengths are derived from industry level marginals. There, the volume (in euros) of a product sold or purchased in a certain industry within a commodity group is given by D_α^{out} and D_α^{in}. The strength of the firm is then calculated as a portion of its net turnover relative to this volume

$$\begin{aligned} s_{i,\alpha}^{out} &= \frac{\text{net turnover firm } i}{\text{total net turnover industry}} D_\alpha^{out} \\ s_{i,\alpha}^{in} &= \frac{\text{net turnover firm } i}{\text{total net turnover industry}} D_\alpha^{in} \end{aligned} \tag{A.1}$$

From these strengths, scores are calculated to assign the most links to the highest scoring firms. Each firm gets a company score $C_{i,\alpha}$ between 0 and 1 given by

$$\begin{aligned} \delta_i^\alpha &= \max_i \left(\log s_{i,\alpha}^{out}\right) - \log s_{i,\alpha}^{out} \\ C_{i,\alpha} &= 1 - \frac{\delta_i^\alpha}{\max_i(\delta_i^\alpha)} \end{aligned} \tag{A.2}$$

The highest scoring firm in a commodity group will have a company score of 1, denoting a high preference from other firms to link to it.

Then a distance score d_{ij} is given, derived from the relative geographical coordinates (x, y) between each pair of firms. It is given by

$$\begin{aligned} \Phi_{ij} &= |x_i - x_j| + |y_i - y_j| \\ d_{ij} &= \frac{\Phi_{ij}}{\max_i \Phi_{ij}} \end{aligned} \tag{A.3}$$

The last score used is the NACE-score I_{ij}. This is an industry score indicating wether or not the using firm's industry actually trades with the supplying firm's industry. Since this is known at industry level from national accounts, the absence of trade between industries to which a given supplier and user belong can be used to give a penalty to such links. It is given as

$$I_{ij} = \begin{cases} 0 & \text{if NACE groups of user } j \text{ and supplier } i \text{ do trade} \\ -1 & \text{else} \end{cases} \tag{A.4}$$

9) Namely the strength/turnover, distance and NACE scores.

A total score is then made by combining the previous scores into an overall score for link-incidence between firm i and j in commodity group α given by

$$\text{score}_{ij}^{\alpha} = \beta C_{i,\alpha} + (1-\beta)(1-d_{ij}) + I_{ij}. \tag{A.5}$$

The β determines the relative importance of each score and can be seen as the tuning parameter in this model. The same value is used across all commodity groups, and chosen to be $\beta = 0.7$.

After the score value, the SN also estimates the number of ingoing and outgoing links through their strengths, i.e. the degrees. The in-degrees are given by

$$k_{i,\alpha}^{in} = \left(\log s_{i,\alpha}^{in} - \min_i \log s_{i,\alpha}^{in} \right)^{\eta} \tag{A.6}$$

The η is assumption-based and chosen to be $\eta = 0.5$. The $k_{i,\alpha}^{in}$ is then rounded down, since degrees are discrete. The out-degrees are then computed via the in-degrees since $\sum_j k_{j,\alpha}^{in} = \sum_i k_{i,\alpha}^{out}$ together with another empirical assumption that a firms degree and its turnover share a powerlaw relation.

$$\begin{aligned} \text{turnover } i &= \Gamma (k_{i,\alpha}^{out})^{\gamma} \\ k_{i,\alpha}^{out} &= \left(\frac{\text{turnover } i}{\Gamma} \right)^{\frac{1}{\gamma}}. \end{aligned} \tag{A.7}$$

The γ is chosen to be 1.3, the resulting powerlaw-relation can be viewed as a parallel to the fitness Ansatz used in the FiCM, as it also embodies an empirical relation between a firms fitness-score and its degrees. The relationship between the degrees and Γ is not continuous and rounding down will sometimes result in degrees[10] of 0. To estimate Γ a bracketing and bisection method is used with an initial estimate:

$$\Gamma_0 = \left(\frac{\sum_i \text{turnover}_i^{(1/\gamma)}}{\sum_i k_i} \right)^{\gamma} \tag{A.8}$$

.

With both the scores and degrees defined, the reconstruction procedure distributes the available connections in the network, ordered via the scores. To be brief, suppose that commodity group α contains N_u users and N_s suppliers. Then, first the users are ordered according to their strength, hence the first user 1 is the one with the highest purchasing volume. This user then has $k_1^{in} = m$ incoming links to distribute, where the suppliers are also ordered according to their scores. Thus suppose firm j has the highest score then $j = \max_j \text{score}_{1j}$. The m highest scoring firms are chosen to make an ingoing connection to firm i and those firms have 1 subtracted from their degree, i.e. $k_j^{out} = k_j^{out} - 1$. Once these links are assigned, the next user is chosen and the procedure is repeated, until all links have been distributed.

Suppose the procedure is carried out for commodity groups $\{1, ..., C\}$ then the algorithmic procedure is given below

[10] Note that the assumption is that the network is connected, thus all nodes have a degree of atleast 1.

Algorithm 3: Deterministic SN-method.

for $\alpha = 1, ..., C$ do
  order users from largest to smallest according to their volume for $j = 1, ..., N_u$ do
    $m = k_j^{in}$ select top $T = \min\{N_s, m\}$ suppliers based on the ordered score where out-degree is at least 1 for $i = 1, ..., T$ do
      Assign link $i \to j$ $k_i^{out} = k_i^{out} - 1$

B Deriving CReM

The degree-correct gravity model may use a probabilistically obtained adjacency matrix but the method in itself is deterministic, since the analytical values of any potential configuration model p_{ij} are used. Furthermore, the possible weights one can assign are usually assumed to be natural numbers. In the derivation of the Directed Weighed Configuration Model (DWCM) this assumption is used to conclude that the weight are geometrically distributed. However, as seen in Parisi et al. (2020), the weights can also be taken as continuous random variables. In doing so one ends up with an exponential distribution for weights. This should be of no surprise, as the exponential distribution is a continuous version of the geometric distribution. Contrary to the DWCM however, it is applied in a 2-step reconstruction procedure. The disentanglement of the DECM allows it to first reconstruct the topological adjacency matrix, and then the weighted configuration. As such, the goal is to find the conditional (on a certain topological configuration $P(A)$) distribution of the weights $Q(W)$ that take the reconstructed links $P(A)$ as input. Whilst the quest for a conditional distributions give rise to some small technicalities, the derivation can be viewed as a continuous derivation of the DWCM.

Start by looking for the conditional probability of the weights, given the adjacency matrix A as $Q(W|A)$ with marginals $q_{ij(w_{ij}|a_{ij})}$. The resulting method is called the Conditional Reconstruction Method(CReM). In order to find the distribution, first reconsider the optimisation of entropy in a continuous setting. Similarly to the discrete setup the following Lagrangian optimisation problem is encountered:

$$\text{maximise} \quad S(\mathcal{W}|\mathcal{A}) = -\sum_{A\in\Omega} P(A) \int_{\mathbb{W}_A} Q(W|A) \log Q(W|A) dW \tag{B.1}$$

$$\text{subject to} \quad \mathbb{E}f_m(W) = \hat{f}_m, \text{ for all } m \tag{B.2}$$

$$1 = \int_{\mathbb{W}_A} Q(W|A) dW$$

As most clearly illustrated by the presence of an integral, the same setup as in section 2.2 in a continuous way. In order to do so, a lot of notation is (re)introduced. Thus, to quickly summarise.

- From earlier S is the Shannon-Entropy, Ω the sample space of the binary configurations A, and f_m the (now continuous) weighted constraints for any m , representing the available information about the network.
- $\mathcal{A}$ and $\mathcal{W}$ are the corresponding random variables of the adjacency and weighted networks.

- The continuous set $\mathbb{W}_A = \{W : \Theta(W) = A\}$ over which is integrated. Here Θ is the binary projection of W onto A.
- It holds that $\sum_{A\in\Omega} P(A) \int_{\mathbb{W}_A} Q(W|A) f_m(W) dW = \mathbb{E} f_m(W)$, and $\hat{f}_m = f_m(\widehat{W})$ with $\widehat{W}$ the true weighted matrix of the network.
- The Lagrangian constraints with multipliers λ_m and μ for all m.

The normalisation of the conditional probability also ensures that the unconditional probability $Q(W)$ is normalized

$$\begin{aligned} \int_W Q(W)dW &= \int_W \sum_{A\in\Omega} Q(W|A)P(A)dW \\ &= \sum_{A\in\Omega} \int_{\mathbb{W}_A} Q(W)dW \\ &= \sum_{A\in\Omega} P(A) = 1 \end{aligned} \tag{B.3}$$

This leads to the following Lagrangian

$$\mathcal{L} = S(\mathcal{W}|\mathcal{A}) + \sum_{A\in\Omega} \mu \left(1 - \int_{\mathbb{W}_A} Q(W|A)dW\right) + \sum_m \lambda_m \left(\hat{f}_m - \mathbb{E} f_m(\mathcal{W})\right). \tag{B.4}$$

Then take a derivative w.r.t. $Q(W|A)$ and solve the root

$$\begin{aligned} \frac{\mathcal{L}}{dQ(W|A)} &= -\sum_{A\in\Omega} P(A)Q(W|A) \log Q(W|A) + \sum_{A\in\Omega} \mu Q(W|A) - \\ &\quad \sum_m \lambda_m f_m(W) \sum_{A\in\Omega} P(A)Q(W|A) \\ \log Q(W|A) &= \sum_{A\in\Omega} \frac{\mu}{P(A)} - \sum_m \lambda_m f_m(W) \\ Q(W|A) &= \frac{e^{-\sum_m \lambda_m f_m(W)}}{e^{-\sum_{A\in\Omega} \mu P(A)}}. \end{aligned} \tag{B.5}$$

Here it is tacitly assumed that

$$q_{ij}(w_{ij}|a_{ij}) \qquad \begin{cases} = 0 & \text{if } a_{ij} = 0 \\ > 0 & \text{if } a_{ij} = 1 \end{cases} \tag{B.6}$$

Note that, with the right μ, the denominator is a normalising constant $Z_{A,\lambda}$ such that it can be denoted as

$$e^{-\sum_{A\in\Omega} \mu P(A)} = Z_{A,\lambda} = \int_{\mathbb{W}_A} e^{-\sum_m \lambda_m f_m(W)} dW. \tag{B.7}$$

Doing so leads to the Exponential random graph form again (with Hamiltonian constraint function $H(W|\lambda) = \sum_{A\in\Omega} \lambda_m f_m(W)$):

$$Q(W|A) = \frac{e^{-H(W|\lambda)}}{Z_{A,\lambda}} \tag{B.8}$$

To continue the analogy with the deterministic method, it remains to figure out a way of choosing the right constraint parameters λ. Earlier this was done via a maximum likelihood argument where the MLE-estimator for the system was exactly the parameter that would enforce the desired constraints. For the continuous set-up something similar is happening in a conditional probability setting. In this setup there is no clearly defined

adjacency matrix A[11]. As such a more comprehensive likehood is required. For this Parisi et al. (2020) uses the (in the introduction abandoned) generalised likelihood $GL(\lambda)$. which is the log-likelihood of $Q(W|A)$ w.r.t. to the unconditional expectation of $\widehat{W}$. The idea is that now it is fine to use a generalised likelihood, as we have full knowledge on the strength degrees conditional on some binary configuration, so no biases are induced on the weights by using it.

$$GL(\lambda) = -H(\mathbb{E}(W)|\lambda) - \sum_{A\in\Omega} P(A) \log Z_{A,\lambda}. \tag{B.9}$$

If the obtained expression of the conditional probability $Q(W|A) = \frac{e^{-H(W|\lambda)}}{Z_{A,\lambda}}$ is plugged into the Lagrangian it follows that

$$\begin{aligned}
\mathcal{L}_{Q_\lambda(W|A)} = S(\mathcal{W}|\mathcal{A})_{Q_\lambda(W|A)} &= -\sum_{A\in\Omega} P(A) \log Q(W|A) \\
&= \sum_{A\in\Omega} P(A) \left(\log Z_{A,\lambda} + \sum_m \lambda_m f_m(\widehat{W}) \right) \\
&= \sum_{A\in\Omega} P(A) \log Z_{A,\lambda} + \sum_m \lambda_m f_m(\widehat{W}) \\
&= -GL(\lambda)
\end{aligned} \tag{B.10}$$

for any $\widehat{W}$ such that our constraints $\mathbb{E}f_m(\mathcal{W}) = f_m(\widehat{W})$ are satisfied. Thus, if $\lambda = \lambda^*$ it would satisfies our desired constraints, where this λ^* will also maximise the general likelihood. Furthermore, mostly using linear constraints are used. Thus, the form of constraints is limited and can consider (for some node i) the Hamiltonian $H(W|\lambda) = \sum_{i=1}^N \sum_{j\neq i} \lambda_{ij} w_{ij}$. The marginals of $Q(W|A)$ are determined as follows

$$\begin{aligned}
Q(W|A) &= \frac{e^{-\sum_{i=1}^N \sum_{j\neq i} \lambda_{ij} w_{ij}}}{Z_{A,\lambda}} \\
&= \prod_{i=1}^N \prod_{j\neq i} \frac{e^{-\lambda_{ij} w_{ij}}}{\int_{\mathbb{W}_A} e^{-\lambda_{ij} w_{ij}} dw_{ij}} \\
&= \prod_{i=1}^N \prod_{j\neq i} \frac{e^{-\lambda_{ij} w_{ij}}}{\left(\int_0^\infty e^{-\lambda_{ij} w_{ij}} dw_{ij}\right)^{a_{ij}}} \\
q_{ij}(w|a_{ij}=1) &= \frac{e^{-\lambda_{ij} w}}{\left(\int_0^\infty e^{-\lambda_{ij} w} dw\right)} \\
&= \lambda_{ij} e^{-\lambda_{ij} w}
\end{aligned} \tag{B.11}$$

The idea behind this specific Hamiltonian is that it can constrain the entire set of expected weights to some given target weights $\tilde{w}_{ij}$. This nets a generalized likehood of the form

$$GL_{CReM} = -\sum_{i=1}^N \sum_{j\neq i} \left(\lambda_{ij} \tilde{w}_{ij} + p_{ij} \log \lambda_{ij} \right) \tag{B.12}$$

where p_{ij} is the binary configuration of choice and maximising it gives

$$\begin{aligned}
\mathbb{E}w_{ij} = \frac{p_{ij}}{\lambda_{ij}} &= \tilde{w}_{ij} \\
\lambda_{ij} &= \frac{p_{ij}}{\tilde{w}_{ij}}
\end{aligned} \tag{B.13}$$

[11] As one wants to be able to use the model for multiple topological reconstructions.

The actual target weights $\tilde{w}_{ij}$ are unknown quantities. In the same vein as the Fitness Ansatz, one can make another ansatz about the relationship between link weights and their node strengths. InParisi et al. (2020) the ansatz of choice is the standard gravity model from section 2.1 for target weights, i.e. $\{w_{ij}^{ME}\}_{i,j=1}^{N}$. The corresponding parameter λ is then

$$\lambda_{ij} = \frac{Wp_{ij}}{s_i^{out} s_j^{in}}. \tag{B.14}$$

where W is the known total weight of the network. The standard gravity model actually performs quite well when only the weights are taken into account. However, this method also allows for the possible incorporation of other weighted configuration like Iterative Proportional Fitting (IPF). Meanwhile, the CReM allows for a probabilistic assignment of continuous link-weights, while only requiring the resolution of $O(N^2)$ decoupled equations.

The Ansatz invoked here is not a necessary component of CReM, one can also use the Hamiltonian $H(W|\lambda) = \sum_{i=1}^{N} \sum_{j \neq i} \left(\lambda_i^{out} s_i^{out} + \lambda_j^{in} s_j^{in}\right)$ and arrive at a similar probability distributions but when deriving the corresponding constraint parameters one needs to solve $2N$ coupled equations. In the reconstruction of large networks this would take significant chunk of extra calculation time, while also combining less nicely with the Fitness model.

Also here, the use of a fitness or rather the target distribution, opens up a possibility of looking at the reconstruction method through a Bayesian lens. While in the CReM case one can choose a deterministic distribution as the target, it could also be another probability distribution to then arrive at a posterior distribution. One can use the Monte-Carlo Markov Chain method or a Gibbs-sampler with this target distribution to arrive at another way of reconstructing links in a probabilistic way. The details for this method will not be discussed but they can be found in Gandy and Veraart (2016, 2019).

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