A methodology for estimating the Dutch interfirm trade network, including a breakdown by commodity

Sjoerd Hooijmaaijers and Gert Buiten[1]

[1] The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

project number	Projectnumber
	sector
	April 4 2019

Index

Abstract

This paper describes the methodology being developed by Statistics Netherlands for estimating the Dutch interfirm trade network using a combination of input data on trade transactions and auxiliary information from existing official statistics and administrative registers. The estimates result in a directed and weighted network dataset, including a breakdown by commodity group derived from the National Accounts' Supply and Use tables. The method is flexible in taking on board various data sources and could be applied by other National Statistical Institutes.

There are various possible sources for collecting Input data on relationships, such as bank transaction data, overviews of debtors and creditors from company administrations and survey data. Each of these sources has some drawbacks and must be complemented by an imputation method for completing the picture. The method developed provides a flexible framework for dealing with a variety of available data sources and can be applied to other countries.
The past decade, a growing body of methods has been developed for estimating interfirm networks. Most of these methods combine macro or meso economic marginals (such as aggregate turnover by industry group) with one or more assumptions on the distribution of variables by company as well as on interfirm connections by company. The paper describes how this can first be expanded upon by using aggregate public data from Supply-Use Tables and Input-Output Tables as auxiliary data, allowing for a matching procedure between suppliers and users by commodity group. In this process, trade relations between companies are imputed if there is no direct information on these relations. Secondly, it describes further improvements based on using non-public micro data on enterprises from various sources, including direct observations of trade relations between companies.

1. Introduction[2]

Economic network analysis at a micro level of firms opens up a range of new possibilities, such as analyses of chain bankruptcies, systemic risk, early warning signals and the spread of ups and downs in the business cycle through the economic web. However, network data at a micro level is only available in a small number of countries. This paper will cover the methodology for estimating an interfirm trade network on a commodity group level, in particular the case of the Netherlands. Both the research project and this paper are a work in progress. In this version, the paper is meant to inform others and to allow for feedback and cooperation.

The methodology combines information from existing official statistics and auxiliary information from a myriad of sources to result in a directed and weighted network. This combination allows for a 'base level' network with minimal input data while leaving room for improvements with each added data source. There are various ways to collect data on actual relationships between companies such as bank transaction data, overviews of debtors and creditors from company administrations and survey data. All of these sources, if available, have drawbacks and must be complemented by an imputation method. By using Supply and Use tables and Input-Output data, we describe a matching procedure between suppliers and users on a commodity group level. This matching procedure takes several variables into account and allows for simple extension with and incorporation of new data sources. Additional information sources allow for further refinements of input and output quality of the network on many different parts. The Structural Business Statistics (SBS) allows for further refinements in which commodity groups are or aren't supplied and/or used by (groups of) companies.

Replicating a network without knowing the network structure is nigh on impossible. The presence of a highly detailed interfirm trade network in Japan (Bernard, Moxnes & Saito (2015)) and Belgium (e.g. Bernard, Dhyne, Magerman, Manova & Moxnes (2018), Dhyne & Duprez (2017) show us some stylized facts which are helpful in various parts of this paper. The presence of power laws in many aspects such as connection distribution (Bernard et al. (2018), Bernard, Moxnes & Saito (2015)), distance distribution (Bernard, Moxnes & Saito (2015), Dhyne & Duprez (2017)) and weight distribution (Bernard et al. (2018)) play an integral part in the estimation of our network. The adding of weights to the directed network allows for unique possibilities in analyzing complex effects of changes in the economy. The current methodology is for one year of data, but can be expanded upon to include multiple years.

[2] The authors wish to thank Patrick Bogaart for his important contributions in the early stages of this project, especially on the basic mechanism of the matching procedure.

2. Terminology and concepts

This section describes the terminology and concepts used in this paper. Starting point is the set of companies available from Statistics Netherlands' Statistical Business Register. With these, we have all the nodes of the network, so the main challenge is to measure or reconstruct the (trade) relations c.q. edges between these nodes. Without any additional information, companies can potentially supply and use all commodities and be connected to all other companies. By adding various types of information and assumptions, we can narrow down the range of possible combinations between companies and between companies and commodities. Companies are part of industries for which the Supply-Use tables tell us which goods categories they can supply and use. These classifications allow us to exclude relationships between companies purely because they do not have a commodity group which they can trade with each other.

The Supply-Use tables alone are not enough to estimate a trade network. For that, we must zoom in to a micro level, as can also be seen in the flow chart of the overall process in figure 1. We create Supply and Use matrices for each industry, containing only the relevant companies and commodities. Additional sources can help us to find specialisation in production or consumption of certain commodity goods. Specialisation in this case can also mean the absence of production or consumption. In section 3.1 we use the Structural Business Statistics as an example to improve accuracy on our micro level Supply-Use matrices. After this step we convert these matrices into commodity group specific matrices, for production and consumption. We then know which companies are possible suppliers and possible users for each commodity group, which defines the pool from which we can match companies together. Since this is the Dutch interfirm network, exports and imports are not taken into account (yet). As such, there may be some commodity groups which are either supplied but not used, or used but not supplied. Neither are included in the estimated network.
Based on the values in our micro level Use matrix we infer the number of ingoing relationships, the number of suppliers from which a firm will purchase a commodity group. We call this the 'indegree'. The total indegree for a commodity group is then distributed over all potential suppliers of this commodity group, such that each company has a number of companies it delivers to. We call this the 'outdegree'. The matching procedure in short uses several characteristics and turns these into scores (see also the flow chart of the matching process in Figure 2). A weighted sum of these scores determine the ranking of all supplying companies of the commodity group. A connection is then imputed between the using company and the top X suppliers, given that the supplying companies have a positive remaining outdegree. With X being the indegree of the using company for this particular commodity group. A more in-depth explanation is given in section 4. Lastly, we distribute the values in our micro level Supply matrix over all the outgoing connections of a company (per commodity group), turning our directed network into a weighted and directed network.

Figure 1: overall process flow chart network reconstruction procedure

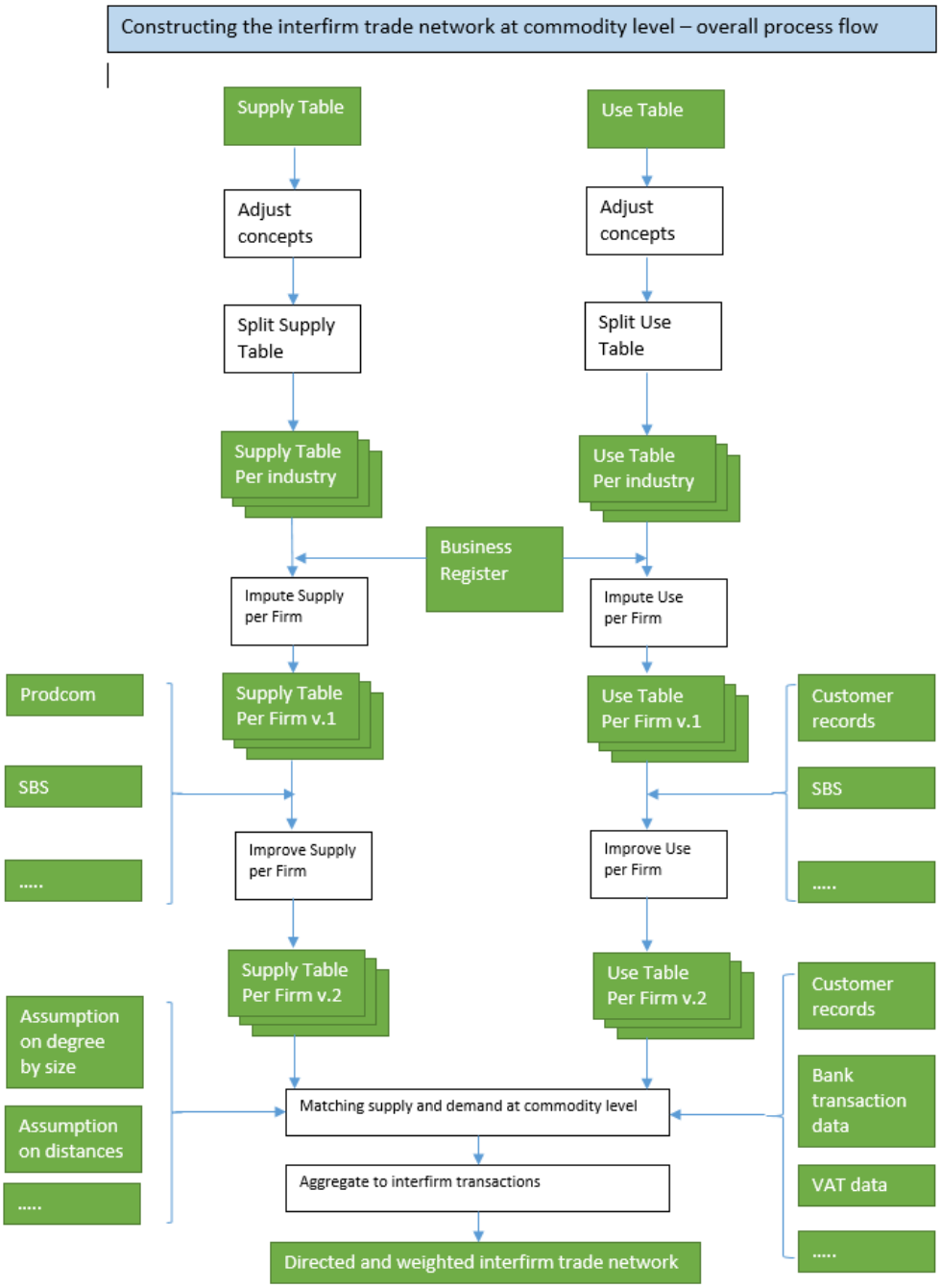
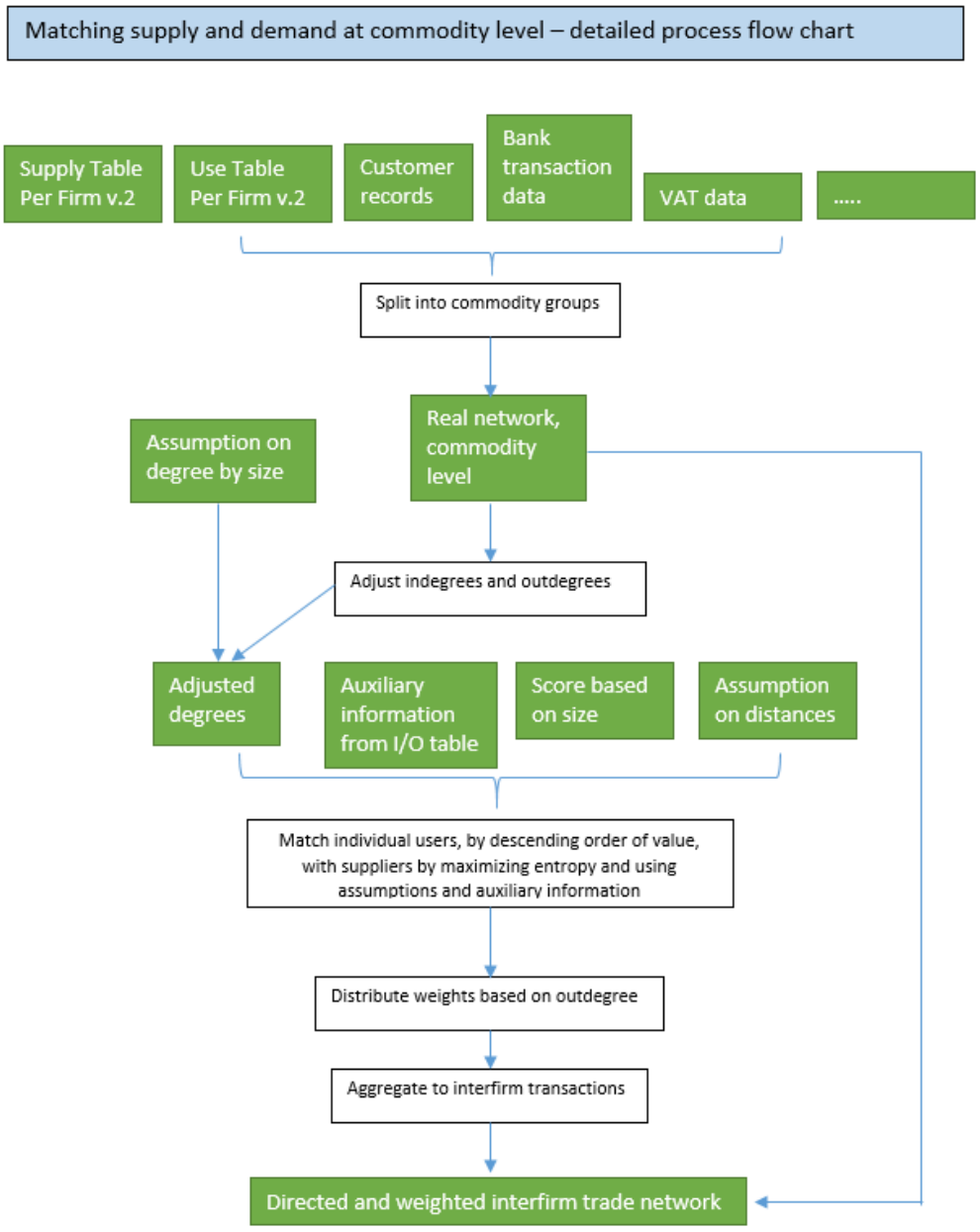


Figure 2: detailed process flow chart matching procedure



3. Creating Supply-Use matrices on a micro level

In order to match supplying and using companies based on goods categories it is imperative to have a Supply-Use table on a more detailed level, that of individual companies. For easier use, we make a Supply-Use matrix for each industry. The dimensions of each matrix are all the companies in the industry by the unique amount of goods either supplied or used. We know, as a given, from the Supply-Use table on a national level what the total value supplied/used by all the companies within a particular industry are for each commodity group. To impute commodity group specific values for each companies, the sum of all commodity groups need to be distributed over all companies in the industry. Without any additional information sources the weight used for this is the net turnover of a company as a fraction of total net turnover in an industry. With these company marginals, which are different from their net turnover, the value per commodity group for each firm is calculated.
There are many (small) companies within an industry, while the amount of different commodity groups can be large (and their value small). It is unlikely that all companies will supply or use all of the commodity groups which the industry does. Although a large company within an industry may specialise in several commodity group, the size of small companies forces them to limit their activities. Due to small company marginals, values for goods in non-primary activities can be unrealistically small. By setting a boundary underneath which values are set to zero, it is possible to implement this assumption into our micro level Supply-Use matrices. The impact of the boundary depends on the industry, mainly on the amount of companies. Further research on a possible optimal value for the boundary and whether this should be industry specific is needed. After this a small step is needed to ensure that the sum of all parts is still equal to the total as calculated before. The cleaning procedure which we use is the iterative proportional fitting (IPF) procedure. This ensures the sum of all rows (firms) and columns (commodity groups) are equal to their respective marginals. In essence, the making of a micro level Supply-Use matrix is not a particularly difficult or time-consuming step if no auxiliary information is available. Any data sources that can help with pinpointing specialisation of companies, or guaranteeing that certain commodity groups are not supplied or used, are a huge step in improving the quality of the network. Incorporating a data source may be non-trivial and lead to potential problems. We will give an example of how to include data sources in the following section by incorporating the SBS.

For one industry an additional action is necessary: in the Supply and Use tables, sales of Wholesale companies to other companies only include sales of goods and services produced by Wholesale companies themselves. The sales of trading goods produced by other domestic or foreign companies are recorded as sales directly from the producers to the using companies, with wholesale trade margins as a separate row (production of wholesale margins by industry) as well as a separate column (wholesale margins by commodity) in the Supply table. In the National Accounts terminology, Wholesale is recorded ‘net’ instead of ‘gross’. This is not in line with the actual trade and payments flows, where Wholesale trade is quite often the intermediary between original producers and the using companies. Therefore, to align these Supply and Use data to actual trade and payment flows a transformation is needed. In the Supply table, the sales of Wholesale companies have to be enlarged with the value of gross wholesale trade flows. In the Use table, the purchases of Wholesale companies for reselling purposes have to be added. The necessary imputation is based on the column in the Supply table containing the total intermediary wholesale trade margins per commodity group. Using a trade margin percentage, the net margins per commodity group can be transformed into total

gross wholesale sales. These totals include sales to other companies, but also to exports and – via Retail trade – to consumers. From this, an estimate of the intermediary Wholesale sales to other companies (excluding Retail trade) can be obtained by using the proportion between intermediary wholesale trade margins and total wholesale trade margins from the Input/Output-table. The figures from this estimate are then added to the column of the Wholesale industry.
Using the same procedure, the purchases for wholesale trade activities per commodity group are calculated and added to the column of the Wholesale industry in the Use table. Currently, one margin percentage is used for all commodities and one fraction for obtaining the intermediary Wholesale sales.

Basically, the same holds for companies in Retail Trade. However in the current first stage of the research this is disregarded since sales from retailers to other companies are relatively small. For the Transport industry something more or less similar holds. Their trade margin is seen as a separate product.

In the current version of our work, we do not use all the companies from the Statistical Business Register, since there are many companies with only small annual turnover figures. To get a clearer picture of the network structure we apply a threshold of an annual turnover of 10.000 euro. This leaves some 870.000 companies (for the year 2012) as nodes in the network with a turnover of at least 10.00 euro. Some 500.000 'micro companies' are left out.

3.1 Structural Business Statistics as auxiliary data source

As an example of how to incorporate auxiliary data sources into the Supply-Use tables on a micro level, we use the Structural Business Statistics (SBS). Since the SBS-survey is by definition a sample, it only covers a small part of all companies. As such, we will also attempt to generalise outcomes if possible. Note that since the value of the marginals remains the same, adding a value to a company for one or more commodity groups with certainty indirectly also gives higher credibility to the values of companies for which we have no additional information. Since the marginals of companies in our Supply-Use matrices differs from their net turnover (which in the Dutch case is taken from the business register), we must transform all information from the SBS into weights.
Companies are not only classified by the industry they work in, but are also sub-classified into Standard Industrial Classification (SIC). Results are attainable on three different levels. Firstly, on the level of SIC. Secondly, on the level of SIC and the same size class. Lastly, we can put the information of the sampled companies within the SBS directly into our micro level supply-use matrices. Now we have to choose a threshold for when the sampled companies in the SBS are representative for their respective SIC or SIC-size class combination. We have chosen a minimum of 10 companies, as well as at least 5% of the companies in the business register. In our case, there are companies who are classified in a different SIC and/or size class in the SBS and the business register. Misclassification in goods categories conversion, as well as companies' SIC are likely a large part of two problems we encountered while incorporating the SBS into our micro level Supply-Use tables.
In order to make a micro level Supply-Use matrix with auxiliary information, we start with a clean slate. The results we find from the SBS are inserted as a percentage of the company marginal. If the sum of all commodity groups for which we have information does not add up to 100% then the remainder is – slightly later in the process – split over the commodity groups for which we have no information. As such, by construction, the sum of all commodity group values for a company is equal to the company marginal. This does not mean that the same holds for

commodity groups. In some extraordinary cases, we come across two distinct problems. One being underfitting, the other overfitting. In the first case, the sum of values from companies we have information (for a specific commodity group) for plus the marginals of all the companies we do not have information for is not enough to sum up to the total value of the commodity group. For the first case we do not yet have a solution and we choose to delete all inputs for this commodity group, within the industry. In the second case the sum of values from companies we have information for already exceeds the total value of the commodity group, even if all values for unknown companies are set to zero. In this case, the value we place for each company is its weight in the industry multiplied by the total of the commodity group. Then, for the companies we have information for we overwrite this 'neutral' value with the value we had. At this point, our cleaning procedure will adjust the values such that the sum is equal to the marginals. This allows companies who said they specialise in certain commodity groups to still do so, however without the unrealistic aspect of forcing all other companies – of which we have no information – to not supply or use this commodity group.

3.2 Imputing the number of in- and outgoing connections

With our Supply-Use matrix on a micro level now completely filled, we need to calculate the amount of in- and outgoing connections of each company: the indegree and outdegree. This is done on a commodity group level. In general, the number of users is larger than the number of suppliers, so we first estimate the total indegree using the Supply matrices. This total is by definition equal to the total outdegree. In a second step we distribute the total outdegree to company x commodity combinations. These are further on in the process used in the matching process. This is explained below.

The calculation of the total indegree is done in a two-step procedure. First, the number of commodity x company combinations in the use matrix is calculated. This is a lower bound for the total indegree, corresponding with the theoretical situation where every using company has only one supplier. Since e.g. branches like Wholesale it is clear that companies may have more than one supplier, in a second step the lower bound is raised with an adjustment factor. This factor is based on an assumption, and may be used for calibrating the method.

By definition, the total indegree is also the total outdegree. This outdegree is distributed over company x commodity combinations. This is done using the findings in the literature that the degree distribution follows a power law distribution related to the size of the company. Watanabe et al. (2013) deduct the following relationship between the turnover S of a company I and its number of connections k:

$$S_i = \alpha k_i^y, \qquad y \approx 1.3 \implies k_i = \left(\frac{S^i}{\alpha}\right)^{\frac{1}{y}}$$

With this, a first estimate of α can be made using:

$$\alpha_0 = \left(\frac{\sum s_i^{\frac{1}{y}}}{k_s}\right)^y$$

Since the relation between α and k is not continuous (k is discrete) and low turnovers may lead to k = 0, an algorithm has been developed which estimates a valid value for α, using bracketing and bisection. Bracketing means that the value for α_0 is estimated as if k would be continuous. Bisections means that from that value an interval is chosen in which the correct value of α should be, after which this interval is made smaller until this correct value is found.

4. The matching process

The matching procedure uses several inputs and variables to create a network on a commodity group level. All commodity group networks are in fact separated from each other. In the matching procedure the using company is connected to a selection of all potential suppliers of a commodity group based on several characteristics. For this, the Supply-Use matrices are transformed from a focus on commodities per industry to firms per commodity. The using company with the highest value for a commodity group has first pick. This company will choose the 'X' companies with which it matches the best, with X being the commodity group-specific indegree for the using company. What is the best match is derived from a score that is calculated with the procedure described in 4.2. The outdegree for the matched (calculated in the step describer in 3.2) supplying companies is adjusted, ensuring that a company with no outdegree remaining can't be chosen. Because of this, as the using companies (in descending order of value) are matched with partners, the pool of possible suppliers is shrinking.
The method to rank all the possible supplying companies consists of several variables and is a method that is open to additional auxiliary sources as well as incorporating real data on relationships between companies. The current method incorporates scores based on the (commodity group specific) size of a company, the (geographic) distance between a user and supplier and a score for certain industry couples based on the I/O table. The implementation of real data requires relationships on a company level to be transformed to a commodity group level. At this point the corresponding indegrees and outdegrees must be adjusted, to keep the total size of the network the same. Likewise, a measure is added to the weighing mechanism to prevent a relationship from occurring both in the real and the imputation part.

4.1 Implementing real network data

Clearly, directly observed data on relations between companies improves the quality of the network estimation. In the Netherlands we aim at using data from company records on customers en suppliers as well as bank transaction data. For direct data on relationships between companies to fit into the process, additional steps are needed. Firstly, we need to decide for which commodity groups we establish a connection in our network. Secondly, after we know which companies are matched for which commodity groups we need to adjust the corresponding indegrees and outdegrees in order for the rest of the network to be imputed properly. Naturally, a relationship can only be established for commodity groups that are supplied/using by the supplying/using company. This already puts a natural limit on the possible number of commodity group relations that are set between two companies. . Establishing a relationship for all these options however, is most likely to be quite the overestimation. The extent of the overestimation depends on the quality and quantity of data used in making the micro level Supply-Use table. More data will lead to more specialisation for (groups of) companies and will decrease the overlap of commodity groups between companies. Our method chooses from the overlap a number depending on the percentile of the supplying and using company within their respective industries. The real data points will be added into the estimated network at a later stage.

4.2 The matching procedure for imputing a network

More often than not, there will not be directly observed data on trade relations for all companies and all commodities. Therefore, the unobserved links of the network must be

imputed or reconstructed. This is done with a procedure that matches suppliers and users per commodity. Per user of a commodity, possible suppliers are ranked by a score that is a measure for the likelihood that they will trade with the using company. This score may be a weighted average of various partial scores, each of which can be based on an assumption and/or on auxiliary information. Once again this provides a generic and flexible approach that can be adapted to the available data. To make scores comparable, they are all expressed expressed as a value between 0 and 1.

4.2.1 Company score

The company score is a score based on the value in our micro level commodity group Supply matrix. To make the scores comparable, all values are turned into a score ranging from 0 to 1. The score is calculated per commodity. The score is relative to all other companies (supplying the commodity group), as such there is always one company with a maximum of 1. We transform the values in their current form by taking the log (base 1). Since log(1) is equal to zero, we calculate the score in the following manner:

$$Logscore_x = \log(net\ turnover_x) - \log(net\ turnover_{max})$$

$$Company\ Score_x = 1 - \frac{Logscore_x}{\max(Logscore)}$$

One of the implicit assumptions here is that large companies will prefer to match with other large companies. Large supplying companies will have a high company score, while large supplying companies will be the first to “choose” their partners.

4.2.2 Distance Score

The distance score is based on the geographical distance between two companies, again scaled from 0 to 1. A growing distance means a larger impediment to a potential relationship for a number of goods where transportation time and costs are increasing with the distance. Bernard, Moxnes and Saito (2015) describe the Japanese production network. They find that geographic proximity plays an important role, with a median distance to a supplier of 30 kilometers while the mean distance is 172 kilometers. They also find that large firms have more suppliers, which are also further away on average. Although the Netherlands is a relatively small country, we expect as in Belgium (Dhyne & Duprez, (2016)) that “even in a small country (…) geographical distance is a key determinant of trade”. The median and mean distance in the Belgian production network are 25 and 38, respectively.
It can be noted that the median distance is surprisingly similar between the perhaps not so similar countries of Japan and Belgium, a finding that could be present in more countries – including the Netherlands. In the network we make, on a commodity group level, it is harder to compare distances between companies to the networks between companies. We can aggregate our commodity group network to a complete network and compare the distribution of the aggregated network to that of Belgium and Japan.
The scaling of 0 to 1 in the intra-national case is logical, as you would compare the transportation time/costs relative to other choices. This does mean that by definition the distribution between 0 and 1 will be different between a central-lying or a company near the border. Since the network is the inter-firm network within a country, neither imports nor exports are taken into account even though foreign companies are substitutes for domestic companies. For each using company we calculate the respective distances between it and all

possible supplying companies. To save on computation time, the score for distance is currently a crude method for comparing the distance relative to the furthest possible partner.
For all possible suppliers the distance score for a using company is calculated as such:

$$Distance_i = abs(Lat_i - Lat_{use}) + abs(Long_i - Long_{use})$$
$$Distance\ Score_i = \frac{Distance_i}{\max(Distance)}$$

4.2.3 Industry score

From the Input-Output table we know values traded between industries on an industry by industry level. There are two ways in which we can use this information to improve the commodity based matching process. First, we add a penalty to the total score if in the Input-Output table two industries do not trade with each other whatsoever – across all commodity groups. Second, we can add a small bonus if two industries do trade with each other.

4.2.4 Weighing of scores

While the industry score is a set bonus (or penalty), we can give differing weights to the company and distance scores per commodity. This allows us to fine-tune the weight we will give to each score. With this, we can try out various weights to see which one results in a more 'realistic' network. We will sum the weights of these scores to 1. As such, the weight of Distance score is equal to 1 – Company score. What will help us with finding a weight is that we can expect a certain distribution of the distances between companies. In Belgium (Dhyne & Duprez, (2016)) there is a thorough dataset based on VAT data. They show a distribution of distances between companies. Although the economies of Belgium and the Netherlands are different, they are also similar and the general shape of this distribution may very well hold. Even if the distribution isn't exactly the same, it gives us an insight into what weights for the company and distance score will be less realistic.
Since all commodity group networks are de facto calculated separately, it is also possible to adjust the weight between distance and company size based on some characteristic. Certain types of goods will be more difficult to transport than others, whereas some services can even be performed online. The overall score per company is calculated using the formula:

$$Overall\ Score = \alpha * Company\ Score + (1 - \alpha) * Distance\ Score + Industry\ Score$$

With Company Score and Distance Score between 0 and 1 and Industry Score either -1, 0 or 0.1.

With these scores, the imputation of missing links is performed per commodity and using company. Per using company, supplying companies are sorted in descending order by the score. For the top X number of firms a connection is imputed, with X being the number of incoming relations of the using company for this commodity (the indegree calculated with as described in 3.2).

4.3 Combining real relations and imputed relations

At this point, we have two data sets containing the (binary) relations between supplying and using companies and the commodities they trade. These are basically lists, that can simply be merged and sorted for the next step.

4.4 Adding weights to the network

At this point we have for each commodity group a directed, but not yet a weighted, network. We know for each company how many outdegrees per commodity group it has. With a large amount of connections, it is unlikely that the weight for all connections is the same. Likewise, it is also difficult to have (strongly) decreasing steps. It is more logical that there are one or several trading partners with which most of the value is traded and many trade partners with which a low value is traded. With this reasoning, we put all the trading partners (of a commodity group) of a supplying company into different groups. In these groups, or 'bins', the size of the bin is inversely related to the value which will be assigned to the bin. Companies within the same bin will have the same weight. This means that across bins, the weight is distributed in weakly decreasing steps.

The amount of companies in a bin is calculated by a power law function, taking into account that the amount of companies in a bin should be increasing. The network was made from the perspective of using companies, but we distribute the value based on supplying companies. This means that if we filter the network to only have the relations for which a specific company is supplying – the size of the using companies it is matched with is ordered from high to low by construction. As such, we will use this ordering to distribute the total value in (weakly) decreasing steps. The value to be distributed is the marginal that of the commodity group for a supplying company in the micro level supply matrix. The total weight is distributed over all the bins by a power law distribution. Within a bin the value for the bin is identically distributed across all companies within the bin. The distribution of values across and within bins helps to amplify the effect of the value distribution.

5. Discussion

5.1 How does our method fit in with other imputation methods?

In recent years, numerous methods have been developed to reconstruct economic networks, especially in the financial world. In (Squartini et al. 2018) an overview is given, in which the various approaches are classified into groups of methods. The approach described above could be seen as an example of an 'exact density method' in the terms of Squartini et al., because we use an explicit assumption on the link density. The basic idea of our imputation method is distributing marginals by maximizing entropy but with adding restrictions based on auxiliary information (e.g. SBS-data, company size) as well as on assumptions inferred from the literature. In doing so, this approach is in line with many other methods. There is a gradual difference with other methods in the amount of auxiliary data that we use, since as a statistical office we have access to a lot of micro data. One aspect of our method is not familiar to most others methods: breaking down the relations between companies to the commodity level.

5.2 Further possible data sources

The application of the method described in this paper uses sources that are available in the Netherlands for use by Statistics Netherlands as the National Statistical Institute. When applying the method for other countries, the choice of data sources can be adapted to available sources in that countries as well as to the possibilities the NSI has to access and process data. In the Netherlands, direct information on relations between firms has to come from company records and from payment transaction data from banks. In some other countries, direct information on relations can also be obtained from VAT-registers. In e.g. Belgium and Estonia companies have to provide the Tax Office not only with data on their sales but also give a breakdown of sales by domestic trade partners. In the Netherlands only export sales have to be specified by trade partner. Other possible direct data sources on relations include: data on debtors and creditors from company records, internet data on 'customer testimonials' on company websites, data on transports et cetera.

Moreover, also indirect information on relations between companies could be used, for instance in the form of an additional score in the matching process. An example could be that firms operating with 'biological' or 'sustainable' production processes get a bigger chance in the matching process for trading with each other than with other companies.

The results of the imputation method can also be improved by adding more specific information on supply and use of companies. In addition to SBS data, also data from Prodcom can be used as well as from more detailed sales records.

5.3 Possible applications, especially for this specific data set

Using the network data set, first of all a number of general network analyses can be performed on the relations between companies, like analyses of the network topology, cluster and supply chain detection, analyses of robustness and fragility of the network for changes in the topology (e.g. because of bankruptcies) as well as for large changes in trade flows. Because of the micro level nature of the data set, also various regional analyses are possible. A variety of examples can be found in the literature, e.g. detection of early warning indicators (Squartini c.s. 2013), detection of clusters (Dhyne and Duprez 2017), detection of production chains (Dhyne c.s. 2015), estimation of risk propagation in networks (Goto c.s. 2017) and analyses of the network topology (Bernard 2018).

Since this data set also includes data on the commodities involved in the relations between companies, additional analyses are possible as well. This is still largely unexplored territory. Possible areas of analysis include analyzing trade flows of specific commodities e.g. energy or waste, analyzing supply chains and dispersion analysis of e.g. higher taxes on the production of specific goods. If additional information is added, also e.g. analyses of CO2 emissions can be performed.

5.4 Further research

The data set as such can – at least in the Netherlands – easily be expanded with the relations of companies with foreign trade partners and relations of firms with employees and owners (using tax and customs records). If access to bank transaction data is possible, also relations with (groups of) consumers can be added, thus partly closing the economic loop.
An interesting possible further expansion is currently being discussed with one of the Dutch major banks: to use weekly aggregates of trade transactions between firms to analyze how the interfirm trade network actually operates in real life, e.g. see which patterns of payments and trade are involved with large sales of companies selling to end users. This also allows to investigate what happens before and after bankruptcies of firms.

Apart from that, further research will be done on validation and subsequent better calibration of the imputation procedure. Four areas of research are currently envisaged. First of all, comparison of the results with results from other countries such as Japan and Belgium. Not so much for the features that are used in the imputation methods such as the relationship between number of links and company size, but for instance for the statistical distribution of distances between companies.
A second research area is a comparison of the results with Dutch data sources that are not used in the imputation procedure. A possible approach could be to apply an (intermediary) turnover growth figure for a number of industries and use the relations to extrapolate growth for other industries and compare the results with actual data.
The third research area consists of a sensitivity analysis of the imputation method. What percentage of the imputed relations emerges more or less regardless of changes in e.g. the total number of links, the number of companies used, various weights of the scores used in the matching process, et cetera? This analysis can be performed on the relations between companies but also on the breakdown into commodities. This type of research should give an indication of the accuracy and usability of a completely or largely imputed data set.
Finally, a fully imputed version of the network can be compared with versions of the network that have partly direct data on relations, with the remainder imputed using the same imputation method. To what extent do ‘real’ relations also emerge from the imputation procedure?

References

Bernard, A. B., Dhyne, E., Magerman, G., Manova, K., & Moxnes, A. (2018). *The Origins of Firm Heterogeneity: A Production Network Approach,* NBER Working Paper, w25441

Bernard, A.B., Moxnes, A. (2018), *Networks and trade*, NBER Working Paper Nr. 24556

Bernard, A.B., Moxnes, A., Saito, Y.U. (2015), *Production networks, geography and firm performance,* NBER Working Paper Nr. 21082

Dhyne, E. and Rubínová, S. (2016), *The supplier network of exporters: Connecting the dots*, NBB working paper Research, nr 296

Dhyne, E., & Duprez, C. (2016*), Three Regions, three economies?,* In: Economic Review of the National Bank of Belgium, December 2016 (iii), 59-73.

Dhyne, E., & Duprez, C. (2017), *It's a Small, Small World... A Guided Tour of the Belgian Production Network,* International productivity monitor, Nr. 32

Goto, H., Takayasu, H. en Takayasu, M. (2017), *Estimating risk propagation between interacting firms on inter-firm complex network*, in: PLOS One, Oct 3 2017

Squartini, T., Lelyveld, I. van, Garlaschelli, D., (2013), *Early-warning signals of topological collapse in interbank networks,* in: Scientific Reports 3:3357 November 2013

Magerman, G., Dhyne, E. and Rubínová, S., (2015). *The Belgian production network 2002-2012*, NBB working paper research nr. 288

Squartini, T., Caldarelli, G., Cimini, G., Gabrielli, A., Garlaschelli, D. (2018), *Reconstruction methods for networks : the case of economic and financial systems*, in: Physics reports 757, 1-47

Tintelnot, F., Kikkawa, A.K., Mogstad, M., Dhyne, E., (2018), *Trade and Domestic Production Networks*, NBB working paper research, No 344

Watanabe, H., Takayasu, H., Takayasu, M. (2013), Relations between allometric scalings and fluctuations in complex systems: The case of Japanese firms', in: Physica A: Statistical Mechanics and its Applications, 392 (4), 741-756, doi:10.1016/j.physa.2012.10.020