Automatic classification of farms and traders in the pig production chain

Lisa Köppel[a,*], Tobias Siems[b], Mareike Fischer[b], Hartmut H. K. Lentz[c]

[a] *CHICAS, Faculty of Health and Medicine, Lancaster University, Lancaster LA1 4YG, United Kingdom*
[b] *Department of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Straße. 47, 17489 Greifswald, Germany*
[c] *Friedrich-Loeffler-Institut, Institute of Epidemiology, Südufer 10, 17493 Greifswald - Insel Riems, Germany*

Abstract

The trade in live pigs is an essential risk factor in the spread of animal diseases. Traders play a key role in the trade network, as they are logistics hubs and responsible for large animal movements. In order to implement targeted control measures in case of a disease outbreak, it is hence strongly advisable to use information about the holding type in the pig production chain. However, in many datasets the types of the producing farms or the fact whether the agent is a trader are unknown.

In this paper we introduce two indices that can be used to identify the position of a producing farm in the pig production chain and more importantly, identify traders. This was realized partially through a novel dynamic programming algorithm. Analyzing the pig trade network in Germany from 2005 to 2007, we demonstrate that our algorithm is very sensitive in detecting traders. Since the methodology can easily be applied to trade networks in other countries with similar infrastructure and legislation, we anticipate its use for augmenting the datasets in further network analyses and targeting control measures. For further usage, we have developed an R package which can be found in the supplementary material to this manuscript.

Keywords: German pork production, Animal movement, Traders, Complex network, Network Analysis, Epidemiology

1. Introduction

The trade in live pigs is an essential risk factor in the spread of animal diseases. This is particularly true for the dissemination of the classical swine fever (Büttner et al., 2013; Fèvre et al., 2006; Fritzemeier, 2000; Green et al., 2006; Ribbens et al., 2011) and also non-notifiable diseases like salmonellosis (Lo Fo Wong et al., 2004) or brucellosis (Daz Aparicio, 2013). In particular in Germany, pork production is a very important economic factor (FMoFaA, 2017). Trade takes place mainly between different agents that contribute to the production chain. These include all different kinds of producing farms such as breeding/nursery/fattening farms, their mixtures,

*Corresponding author
Email addresses: koeppel.lisa@gmail.com (Lisa Köppel), tobias.siems@yahoo.com (Tobias Siems), email@mareikefischer.de (Mareike Fischer), hartmut.lentz@fli.de (Hartmut H. K. Lentz)

slaughterhouses and traders. The set of all involved holdings and their trade connections form a complex trade network that can be thoroughly analyzed by Social Network Analysis (SNA) (Dubé et al., 2009; Martínez-López et al., 2009). This has already been performed for the German pig trade by Büttner et al. (2013); Koher et al. (2016); Lentz et al. (2016, 2011) and also for the pig trade in other countries by Bigras-Poulin et al. (2007); Nöremark et al. (2011); Rautureau et al. (2012). In the afore mentioned approaches, the major focus was the ranking of the holdings, which then can be used for optimized control strategies. These rankings can be determined efficiently through methods of network analysis.

The animal movement data typically used for network analyses are collected centrally in the respective countries, and they differ in quality. In Germany, for instance, every single agent in the pig production chain as well as trading partners are recorded. Although a classification about the agents' holdings types is available, it can be misleading, since the farm types are not standardized and moreover, multiple types can be given to each farm. As a result, in network analyses traders might be classified as fattening farms despite the fact that they display strong varying patterns in their trading behavior. However, in order to develop optimized control measures for the case of a disease outbreak, it is strongly advisable to explicitly take the production type of each one of the agents into consideration. A farm that is situated at the beginning of the production chain is for example little suited for being a sentinel for disease detection.

Apart from producing farms, traders play a key role in the trade network, because they often solely liaise between the other pig holdings. Hence, they are involved in a large amount of pigs moved and by being logistics hubs, traders can be crucial in the spread of infectious diseases. Apart from that, when simulating outbreaks on a trade network, traders distort the outbreak dynamics in a severe manner. For this reasons, they have to be distinguished from producing hubs like fattening farms.

The purpose of this study was to find a way of classifying the types of agents in a trade network only through their trading behavior. It included the study of possible holding types, the display of four different trading patterns, and the identification of the position (at the beginning, in the middle or at the end) of a holding within the production chain. With its help, datasets that do not contain the information a priori can be augmented. The focus of our research lied on the detection of traders due to their central role in the trade network. Hence, we developed criteria which allowed us to distinguish between producing farms and traders by analyzing their trading behavior. This was accomplished through a dynamic programming algorithm whose mathematical background and concept can be read in the supplementary material of this article. Our algorithm has been implemented in R and is publicly available from the supplement of this manuscript. In doing so, we achieved a classification of the holdings in the dataset regarding their type in the pig production chain.

2. Materials and Methods

2.1. Data

The dataset used was obtained from the German federal database for livestock movements, HIT (StMELF, 2012). This data storage was established after the BSE crisis in Europe in 2001 to keep record of all trading activities between holdings in Germany. With its help, potential pathways of disease spreading through livestock trade can be detected.

When a pig trade takes place, the receiving holding is obliged to enter the information of the transaction into the HIT database within 14 days of purchase. Here and throughout the manuscript,

with transactions we solely refer to animal movements between two holdings in contrast to financial transactions. The record comprises information about the source and destination premises, the number of pigs moved and the date of the transaction. The temporal resolution is 1 day, thus it is possible to have several records involving the same premises and the same date, meaning the same day of transaction. Because the data is entered by the buyer, the dataset is lacking information about animals moved abroad. Therefore, these records are *not* visible in our dataset.

The analyzed data covered the period from 1 January 2005 to 31 December 2007. In total, 5,528,940 animal movement records involving 70,735 agents were identified. The terms premises, holdings and agents in the context of this manuscript refer to all contributers of the pig movement network,e.g. pig producing units and traders. We only took premises into consideration with more than 10 data records within the observation period; less activity would not suffice to classify the holding. By doing so, we discarded 54,930 premises (43.7%), which were however involved in only 3.51% of all animal movements, representing 1% of the total traded pig volume (3,875,745 out of 361,203,528 traded pigs).

It should be noted that the pig trade network is based on a production chain, which is displayed schematically in Figure 1. The different stages within the production are determined through weight limits of the pigs that are specific to each production step. Hence, in one transaction a group of pigs that roughly have the same weight are being moved from one farm to another. In doing so, livestock can be traded between the holdings directly or by the help of a trader, which is indicated by a T on the edges in Figure 1, which represent animal movements.

[Figure 1 about here.]

The different holding types of the production chain are not available in a sufficient quality for further analyses on the trade network. Hence, the objective of this study was to infer the holding type for each of the premises by analyzing the corresponding trading behavior displayed in the data. It is important to emphasize that we strove for a slightly looser classification than given in Figure 1. This is reflected by the fact that we combined different holding types into one class, e.g. breeding farms (B) were clustered together with breeding-like farms, which included breeding-nursery farms (BR) among others. This also implies that we did not distinguish between fattening and nursery farms or a mixture of the two as they display the same trading structure.

In the pig trade network, the traders play a significant role since they are responsible for movements of large numbers of pigs. Their task involves the logistics of pig transports between the pig producing farms. In this paper, we introduce the reader to methods which allowed us to classify traders and other holding types using the given data.

In order to get a first impression of different trading behavioral patterns, we investigated the cumulative trading volume over time for each of the premises, i.e. at time $t = 0$ the cumulative number of traded pigs was 0 and depending on the respective trading transactions of a holding (purchases or sales), the number of purchased/sold pigs was added/deducted. Examples of four premises are displayed in Figure 2.

[Figure 2 about here.]

Figure 2 a) shows a trading behavior which indicates a breeding farm. The considered holding sold about 7,000 pigs within the observation period. Compared with the sales, the data show only some small purchases (small spikes in the curve progression) from time to time, which can be interpreted as purchases of replacement gilts.

In contrast, Figure 2 b) displays a trading behavior which is characterized through a highly pronounced number of purchases. According to the data the premises did not sell any pigs, which is why we assume that the trading behavior reflects a slaughterhouse.

Figure 2 c) shows a holding with an oscillating behavior of purchases and sales. This approximate balance between purchases and sales signifies that purchased pigs are sold again after some time. Such a pattern is common for holdings which are located in the middle of the production chain, i.e. nursery or fattening farms.

Figure 2 d) displays the trading behavior of a trader. It is characterized by the fact that after a purchase, pigs were sold again within a very short period of time (two days). This is represented by the very high frequency of peaks or vertical lines in the figure. Compared to fattening farms in Figure 2 c), the frequency of the peaks is much higher. It should be noted that the constant increase of the cumulative volume in Figure 2 d) is presumably due to sales abroad, which were not included in the data. And thus, despite the lack of data, we were still able to clearly attribute this plot to a trader.

2.2. Criteria to distinguish between agent types

We sought a classification of the holdings into groups with similar trading behavioral patterns (see Figure 2). Given solely the animal movement data, we introduced two indicators: (1) the *Purchases-Sales-Balance*, which gives information about the location of a holding within the production chain, and (2) the *Trader Index*, which can be used to distinguish traders from fattening farms.

2.2.1. Purchase-Sales-Balance

Given the animal movement data, we denoted the total purchasing and sales trading volume of a single agent over the whole period of the observation with P and S, respectively. We then defined the *Purchases-Sales-Balance* $\boldsymbol{B}$ for a holding through

$$\boldsymbol{B} = \frac{P - S}{P + S} \tag{1}$$

and refer to it as the *Balance*. The Balance corresponds to the net purchases of a holding throughout the entire period normalized by the total trading volume of this holding.

In principle, the Balance represents the ratio P/S between purchase and sales volume of an agent. However, since this is unbounded and not symmetric, we chose Definition 1 to gain values only between -1 and 1. If $\boldsymbol{B} = 0$ holds for a holding, its purchasing volume and its sales volume are equally balanced. A holding with $\boldsymbol{B} = -1$ solely sales, whereas with $\boldsymbol{B} = 1$, it exclusively purchases. In other words, the balance tells us whether an agent's trading behavior was heavy on purchase or sales volume, or was equally balanced, which can be interpreted as an agent lying at the beginning, at the end or in the middle of the pig production chain respectively.

2.2.2. Trader Index

Traders and fattening farms usually resell all the pigs they have purchased and hence their Balance is similar and lies approximately at 0. This means that the Balance could not be used to further distinguish between these two agent types. Therefore, we have developed a second quantity, which is an index that deals with the trading frequency of a holding.

In contrast to fattening farms, who keep their pigs for a longer period on their site, traders always sell their purchased animals shortly afterwards, because they solely convey the animals

in their transporter. According to the Council Regulation (EC) No 1/2005, the duration of a national transport of pigs must not exceed 24 hours (European Union, 2004). Hence, the animals are transported from the seller to the buyer either on the same day, or in case of an overnight transport, at the latest on the following day. Looking at the trading behavior of a trader in Figure 2, this becomes graphically apparent by very small peaks or vertical lines. Each such peak represents a purchase followed immediately by a sale.

From our dataset, for each holding we knew the time and especially the pig volume of every purchase and sale. In the following we consider a purchase to be *prompt* or sold *promptly* if it took at most two days, namely the same day or the day after, to get all of its pigs sold again.

In order to classify a trader, we could have simply checked whether all of his purchases were sold promptly. However, the information given in our dataset did not allow us to discern in which precise transaction a particular batch was purchased/sold. Hence, we could not infer directly when the same batch was traded and moreover, if a purchase was sold promptly or not. In order to circumvent this problem, we have developed a quantity with respect to a single agent which theoretically allowed us to declare as many purchases as possible to be sold promptly on the basis of his trading behavior. This maximization ensured that the true number of prompt purchases was not underestimated and thus, traders could be found with a high sensitivity, at the cost of a lower specificity. Nevertheless, the accuracy was still affected by missing/false data and pig mortality. We addressed this issue by considering agents that had a high number of prompt purchases as traders and not only those where every purchase could be declared as prompt.

In order to develop such a quantity, in the following we consider the trading behavior of an arbitrary but fixed single farm. Let ℓ denote the number of purchases (number of transactions and not the pig volume) of this farm. For each of these ℓ purchases we do not know from the dataset whether they were sold promptly or not, meaning which of the batches were sold within two days of purchase or whether they had been kept longer on the farm and were sold later. However, we can assume one theoretical possibility by considering a function f which indicates whether a purchase indexed by $\{1, ..., l\}$ is either sold promptly or not. That is, if the i-th purchase is sold promptly, then $f(i) = 1$, else $f(i) = 0$. From now on we refer to such a function as an *assignment*.

We can count the number of prompt purchases through $f(1) + f(2) + \ldots + f(\ell)$. To this end, we define the so-called *Trader Index of an assignment* f as

$$\boldsymbol{I}^f = \frac{1}{\ell}\Big(f(1) + f(2) + \ldots + f(\ell)\Big).$$

The Trader Index is a relative number which expresses what fraction of purchases was sold promptly, and therefore lies between 0 and 1. It allowed us to compare different agents, even if they bought and sold at different scales, i.e. they had varying amounts of transactions.

Since prompt purchases are characteristic of a trader, we wanted to find as many as possible using the incomplete information provided in the dataset. Therefore, in theory we had to consider all possible combinations of purchases being prompt or not in order to find the assignment that lead to the maximum amount of prompt purchases and thus results in the largest value for its trader index.

[Table 1 about here.]

However, when considering all possibilities, different assignments may vary in regard to their ability to explain the observed trading patterns. We did not want to consider assignments that

theoretically existed but could not arise in practice. These assignments could be neglected by means of the novel concept of validity. To be more precise, an assignment is called *valid* if the amount of pigs sold by prompt purchases is not higher than visible in the data. It is to note that here we allowed that a purchase could be split and sold in smaller portions, which was reasonable as the pigs were sold depending on their individual weight. An example will motivate this concept further.

For illustration purposes, we now consider an An excerpt of an exemplary farms trading pattern (Table 1). On day t, there are two purchases, indexed i and $i+1$, with 2 and 6 pigs and a sales volume of 3 pigs in total. On day $t+1$, there is one purchase, $i+2$, with 6 pigs and a sales volume of 5 in total. Furthermore, there are no purchases and sales on day $t+2$. Considering all different combinations of prompt and not prompt purchases, with 3 recorded purchases there exist 8 different assignments. However, according to the data, an assignment f can only be valid if it assigns zero to the purchase $i+2$. This is because a holding is not able to sell 6 pigs on day $t+1$, while there are only 5 sold pigs registered in the data. Since it is possible to split purchases, we can at most sell both purchases from day t without breaching the validity constraint: $2+6 \leq 3+5$. The highest possible number of purchases that can be declared as prompt here is obviously 3, however, by meeting the validity constraint we can reach at most 2.

When considering the full dataset, we have to consider the carry of day $t1$ as well in order to ensure validity: If there was another prompt purchase on day $t-1$ which had not been sold completely on day $t-1$, then the remaining amount of pigs had to be sold on day t. By this, a chosen assignment could become invalid. This shows that validity is a statement which incorporates the whole dataset simultaneously. In the supplementary material we addressed this difficulty by means of a recursive formulation of validity.

Incorporating the above, we seek for the *highest possible Trader Index* $\boldsymbol{I}$ under all valid assignments:

$$\boldsymbol{I} = \max \left\{ \boldsymbol{I}^f \;\middle|\; f \text{ is a valid assignment} \right\}. \tag{2}$$

Hence, $\boldsymbol{I}$ can be considered as the highest fraction of prompt purchases that can arise in practice. Throughout this manuscript we will from now on refer to $\boldsymbol{I}$ as the *Trader Index*.

Finding $\boldsymbol{I}$ is an optimization problem, and because the number of possible assignments f is 2^ℓ, we require an efficient algorithm to be able to solve it. In B of the supplementary material, we describe an algorithm to compute $\boldsymbol{I}$ which is easy to implement and runs in polynomial time. Complementary material A deals with the mathematical background of this approach, in which we properly define what validity means, introduce a dynamical programming algorithm and prove its correctness.

From the definition of $\boldsymbol{I}$, we would expect that it separates traders with a high Trader Index from non-traders with a low Trader Index. At this point it should be noted that breeding farms can also obtain high values for their trader indices: now and then they purchase a small amount of breeding sows, but sell their piglets very frequently. In the data, such a purchase with a following (piglets) sale can wrongly be interpreted to be a prompt purchase. However, we will address this issue by means of the Balance, which will classify these agents as breeding farms. In contrast, a fattening farm is very unlikely to have a high Trader Index unless it trades extremely intensively.

2.3. Classification by sight

Unfortunately only very little census data, which is also not specific enough, are currently available for validating the placement of the thresholds. However, for the reason of evaluating the algorithmic findings nonetheless, we applied a different method of classifying the farms which we will refer to as classifying by sight. As displayed in Figure 2, there are specific patterns in a trading behavior of a farm which we took as a standard for each group. In a blind study, we took a sample of 2,000 random holdings out of the dataset without knowing their algorithmic classification. To each one of the 2,000 cumulative plots we tried to assign one of the four groups, namely breeding farm, nursery farm, slaughterhouse, or trader, solely by their characteristics described in Section 2.1. However, if a trading pattern of a farm did not match any of the described types or in case of uncertainty, we labeled the plot as "unknown".

We want to emphasize that absolutely no expert knowledge or background information of the trading business in Germany was needed here in order to assign the according group to a cumulative plot. The characteristics describing the feature of every group from Figure 2 were mainly very clearly visible and easy to spot. We supposed the sample size to be large enough as the relative proportion of each group has reached an approximately steady value while increasing the number of holdings.

[Figure 3 about here.]

3. Results

Figure 3 a) displays the histogram of the Balance of the agents in our dataset. As expected, three modes could be identified. This justifies that the balance was a reasonable quantity to fathom the position of an agent within the pig production chain, albeit not sufficient for a complete classification: traders and fattening farms both got a Balance around 0.

Figure 3 b) shows the histogram of the Trader Index. As anticipated, $\boldsymbol{I}$ mainly divided the agents into two classes: traders with a high Trader Index around 1, and non-traders with a lower Trader Index. Just like the Balance, the Trader Index alone was not sufficient for a complete classification of holdings. To this end, the desired classification was gained by combining these two quantities.

[Figure 4 about here.]

In Figure 4, a scatter plot of the Balance and the Trader Index is displayed. We could clearly identify four point clouds in the picture that are highlighted by different colors. Using certain thresholds, we were able to categorize all agents of the dataset as follows:

- Breeding farms $\mathcal{B} = \{\text{Agents with } \boldsymbol{B} \leq T_1\}$,
- Fattening farms $\mathcal{F} = \{\text{Agents with } \boldsymbol{B} \geq T_1, \boldsymbol{B} \leq T_3 \text{ and } \boldsymbol{I} \leq T_2\}$,
- Traders $\mathcal{T} = \{\text{Agents with } \boldsymbol{B} \geq T_1 \text{ and } \boldsymbol{I} \geq T_2\}$,
- Slaughterhouses $\mathcal{S} = \{\text{Agents with } \boldsymbol{B} \geq T_3 \text{ and } \boldsymbol{I} \leq T_2\}$,

whereby we set the values of the thresholds to $T_1 = -0.5, T_2 = 0.7$ and $T_3 = 0.7$. Our strategy to derive these thresholds was based on an analysis of the trade between the four holding types defined above. This is further discussed in C of the supplementary material.

Our focus was to identify traders in the data, and hence the classification for the producing farms was assessed to be more loose than indicated by the class names. We refrained from further distinction of the producing farms, which would include e.g. mixed holdings amongst others. Instead, their classification aimed at providing an indication of the location of the agents in the pig production chain.

The analyzed dataset did not comprise any information concerning the holdings' true classification. Yet, in order to review the meaningfulness of our thresholds and to verify our algorithmic approach, we used the method of classification by sight. Neglecting the 174 holdings that were categorized "unknown" (8.7%) by sight, we compared the results of the remaining 1,826 holdings with the algorithmic classification.

[Table 2 about here.]

In Table 2, a comparison of the results of the algorithmic approach and the classification by sight is displayed. It became clear that when only considering the sample of the 1,826 holdings as described above, the proportion of each type in both classification methods were approximately the same. Whereas, the fractions of slaughterhouses was virtually the same, the proportions of the other three group types differed only slightly. Regarding the whole dataset with all 70,735 agents, the algorithmic approach lead to a bit more varying proportions of the group sizes compared to the results of the smaller, but assumingly truely classified subset. Notwithstanding, taking into account that the whole dataset comprised wrong and missing data, the algorithmically inferred proportions regarding all 70,735 agents was considerably close to the results of the subset.

Furthermore, by looking at the 1,826 holdings classified by sight, the algorithmic approach matched in 1,786 cases. This resulted in an overall sensitivity of the algorithmic approach (the test's ability to correctly detect the types as they were classified by sight) of 97.8%. Besides, we achieved high sensitivity for the individual classes, meaning the portion of holdings that were classified equally by both methods: for $\mathcal{B}$: 97.9%, $\mathcal{F}$: 98.3% and $\mathcal{S}$: 97.8%. Analogously, the specificity (the test's ability to correctly reject plots for a specific group, that were classified by sight to belong to a different group): $\mathcal{B}$: 96.2% and $\mathcal{F}$: 88.8% and $\mathcal{S}$: 99.0%.

The sample size with 12 traders identified by sight is too small to calculate these quantities for the group of traders. However, since identifying traders was one of our major concerns, we took a slightly different approach for validation. First we assumed that holdings with $\boldsymbol{I} \leq 0.05$ could not be traders, because the Trader Index already returns the maximal relative number of promptly sold purchases which could be inferred from the data. Therefore, we concluded that with less than 5 % of promptly sold purchases the holding was unlikely to be a trader. Analogously, premises with $\boldsymbol{B} \leq -0.95$ were breeding farms due to the fact that in contrast to sales we could see all purchases in the data and nevertheless their sales prevailed by far. In total, there remained only 7,826 holdings that lied somewhere in the middle of the scatter plot in Figure 4.

Out of these specified holdings we took a sample of randomly selected 1,000 premises, roughly about 13%, and classified them in a blind study by sight either to be a trader or not as described in Section 2.3. As a result we gained a sensitivity of 91.9% (68 out of 74) and a specificity of 95.9%.

4. Discussion

In this paper we analyzed the pig trade network in Germany in the timespan from 2005 to 2007. The observed data comprised information about trading activities, but no further information was given about the varying trading agents themselves. However, knowing the position of a producing

farm within the pig production chain and, more importantly, whether the agent is a trader, is of great benefit when analyzing the underlying network. Hence, the objective of this manuscript was to classify the agents in the trade network into one of the four types: breeding farms, fattening farms, slaughterhouses or traders.

Our major focus in this paper lied on classifying traders because they are only responsible for the logistics and the transport of pigs. In case of an epidemic outbreak this can become a significant risk factor: through a large number of traded pigs and a high frequency of transactions, traders might distribute the disease to a wide range of different farms and hence multiply the number of outbreak sources.

Moreover, to balance supply and demand, traders can have several trading partners all across the country and abroad. Trading over wide distances may result in a geographically larger distribution of the disease, which makes it more difficult to minimize the outbreak. Therefore, being able to identify traders in a network can lead to a deeper understanding of the trading structure, which helps to predict potential pathways of the disease spread and to minimize the outbreak.

Apart from being a risk factor, traders could also help indicating a structural change in the pig trade network. If trading partners are not available due to imposed restrictions by the government (e.g. culling of farms, trading restriction zones), traders might take their place and serve as temporary trading partners during this transition period. Hence, they could be used to measure and comprehend such varying dynamics of the different pig producing farms.

For an automatic classification of the agents in the dataset with regard to their holding type, we developed two indices: the Balance and the Trader Index, which are both based on the trading activity of a single agent. Whereas the Balance clarifies the position within the pig production chain, the newly developed Trader Index identifies agents with a very high frequency of selling shortly after buying.

Combining the two index numbers, we were able to assign a type to each of the holdings by setting certain thresholds. Our method is based mainly on unsupervised learning because we were lacking the true classification of the holdings in the dataset. Moreover, sufficient to-date public census data was not available to compare our results to.

Thus, for the purpose of verification we instead used the method of classification by sight. Taking this as the true classification, our algorithmic approach led to very high values in all four class types regarding sensitivity and specificity. In fact, out of a sample of 1,826 randomly chosen agents almost 98% were classified correctly by our algorithm. This means that we were very precise with assigning the right type of $\mathcal{B}, \mathcal{F}, \mathcal{S}, \mathcal{T}$ to a holding just by considering its trading behavior. Moreover, we were very successful in sorting out non-traders as indicated by the high specificity value.

Here, the assessments gathered by classifications by sight were used as a reference and reliable method in finding the true types of the holdings. In a blind study, randomly chosen cumulative plots were analyzed regarding different characteristics that were typical for each class. These included the display of longterm trends in the cumulative trade volume, regular "up-and-down" variations in the curve, and the appearance of spikes, which were vertical lines in the curve progression. With only the help of these descriptions we were able to clearly identify 91.7% of the classified holdings. The remaining 8.3% were too deficient for a sound assessment.

Following these observations, the algorithmic approach was developed to mark the occurrence of trade spikes (Trader Index) and to assign numbers to trends in the cumulative trading volume (Balance). Here, we want to stress again that it would not make any difference whether the person classifying by sight has insight on the algorithmic approach or not. However, as our results showed,

the difference between the two methods came into place when considering the overall quality of the data, which is rather poor. In contrast to using the algorithmic approach, by sight it was possible to handle wrong information such as a sudden purchase/sale of thousands of pigs at once, which clearly affected the Balance. Further, missing information like the lack of purchases, e.g. due to sales abroad (see Figure 2 d)) or the culling during an outbreak, or also structural changes in a farm influenced the outcome of the index numbers.

It should be noted that due to the lack of reference data for the classification we have to make the strong assumption that the by sight classification gives the correct node type. Even though this might not always be the case, classification by sight could in this case be considered a reliable method to serve as a reference for the true classification as it neglected outliers and wrong data and, as we have shown, was able to clearly identify most of the holdings. Thus, high sensitivity and specificity values towards our reference, confirm the credibility of this algorithmic approach.

Since our focus lied on detecting traders, our method did not distinguish explicitly between all different kinds of holdings, for instance mixing types, as our groups were chosen more loosely and group different types together. Nevertheless, for the producing farms our classification approach provided further insight into the placement within the production chain.

A limitation of our method is to distinctly discriminate traders from very large fattening operations. A single holding with multiple fattening barns can display similar structure to a trader as it can have large volume and frequency of pig movements onto site, and large movements off-site. Here, we want to point out that the pig trade in Germany is rather divided among several smaller agents than carried out by monopolists. Hence, fattening farms with relatively many prompt purchases are rather rare. Nevertheless, the threshold T_3, which separates traders from fattening farms, was set such that it is in favor of traders. This means that we took into consideration that we might have classified fattening farms as traders for the benefit of classifying traders correctly, which lead to a high sensitivity in detecting traders at the cost of a low specificity. In terms of classifying holdings only through their trading behavior, it might be sensible to group these two into the same class. Notwithstanding, for a further investigation of the discrimination of the two classes, certainly more information in the dataset would be necessary. Here, sufficient public census data would be of great benefit and would give rise to further research. In order to be able to classify an agent, more than 10 transactions within the observation period had to be present in the dataset. Hence, considering longer time periods might extent the classification to holdings that do not trade much.

The respective R package to calculate the index numbers of our method is publicly available in the supplementary material to this manuscript. An augmentation of a pig trade dataset with our results can be of great use for future network analysis. Concerning the temporal evolution of the node classification, it has been shown that the considered network remains relatively stable over time (Lentz et al., 2016). Thus, it is reasonable to assume that the nodes' roles remain constant as well. In addition, it would not make much sense, if for instance a farm changes from being a trader in one year to being a fattening farm in another year.

This leads to the conclusion that in combining the two indices, our method reached a promising classification of the holdings in the pig production chain.

Acknowledgements

The authors gratefully thank the Friedrich-Loeffler-Institute Riems, Federal Research Institute for Animal Health, for supplying all data used in this study. Furthermore, in memory of Martin

Riedel, the authors appreciated very much his help in implementing the graphical tool for a fast classification by sight.

List of Figures

References

Bigras-Poulin, M., Barfod, K., Mortensen, S., & Greiner, M. (2007). Relationship of trade patterns of the danish swine industry animal movements network to potential disease spread. *Preventive veterinary medicine*, *80*, 143–165.

Büttner, K., Krieter, J., Traulsen, A., & Traulsen, I. (2013). Static network analysis of a pork supply chain in northern germany—characterisation of the potential spread of infectious diseases via animal movements. *Preventive veterinary medicine*, *110*, 418–428.

Dubé, C., Ribble, C., Kelton, D., & Mcnab, B. (2009). A review of network analysis terminology and its application to foot-and-mouth disease modelling and policy development. *Transboundary and Emerging Diseases*, *56*, 73–85.

Daz Aparicio, E. (2013). Epidemiology of brucellosis in domestic animalscaused by brucella melitensis, brucella suis andbrucella abortus. *Rev. sci. tech. Off. int. Epiz.*, *32*, 53–60.

European Union (2004). Council regulation (ec) no 1/2005. *Official Journal of the European Union*, *48*, Chapter V, 1.4.

Fèvre, E. M., Bronsvoort, B. M. d. C., Hamilton, K. A., & Cleaveland, S. (2006). Animal movements and the spread of infectious diseases. *Trends in Microbiology*, *14*, 125–131.

FMoFaA (2017). Understanding farming - facts and figures about german farming. URL: `http://www.bmel.de/SharedDocs/Downloads/EN/Publications/UnderstandingFarming.html`.

Fritzemeier, J. (2000). Epidemiology of classical swine fever in germany in the 1990s. *Veterinary Microbiology*, *77*, 29–41.

Green, D. M., Kiss, I. Z., & Kao, R. R. (2006). Modelling the initial spread of foot-and-mouth disease through animal movements. *Proceedings of the Royal Society B: Biological Sciences*, *273*, 2729–2735.

Koher, A., Lentz, H. H. K., Hövel, P., & Sokolov, I. M. (2016). Infections on temporal networks—a matrix-based approach. *PLoS one*, *11*.

Lentz, H. H. K., Koher, A., Hövel, P., Gethmann, J., Sauter-Louis, C., Selhorst, T., & Conraths, F. J. (2016). Disease spread through animal movements: A static and temporal network analysis of pig trade in germany. *PLoS one*, *11*.

Lentz, H. H. K., Konschake, M., Teske, K., Kasper, M., Rother, B., Carmanns, R., Petersen, B., Conraths, F. J., & Selhorst, T. (2011). Trade communities and their spatial patterns in the german pork production network. *Preventive veterinary medicine*, .

Lo Fo Wong, D., Dahl, J., Stege, H., van der Wolf, P., Leontides, L., von Altrock, A., & Thorberg, B. (2004). Herd-level risk factors for subclinical salmonella infection in european finishing-pig herds. *Preventive veterinary medicine*, *62*, 253–266.

Martínez-López, B., Perez, A. M., & Sánchez-Vizcaíno, J. M. (2009). Social network analysis. *Review of general concepts and use in preventive veterinary medicine. Transbound Emerg Dis*, *56*.

Nöremark, M., Håkansson, N., Lewerin, S. S., Lindberg, A., & Jonsson, A. (2011). Network analysis of cattle and pig movements in sweden: Measures relevant for disease control and risk based surveillance. *Preventive veterinary medicine*, *99*, 78–90.

Rautureau, S., Dufour, B., & Durand, B. (2012). Structural vulnerability of the french swine industry trade network to the spread of infectious diseases. *animal*, *6*, 1152–1162.

Ribbens, S., Dewulf, J., Koenen, F., Laevens, H., & de Kruif, A. (2011). Transmission of classical swine fever. a review. *Veterinary Quarterly*, (pp. 146–155).

StMELF (2012). Herkunftssicherungs- und informationssystem für Tiere. URL: `http://www.hi-tier.de`.

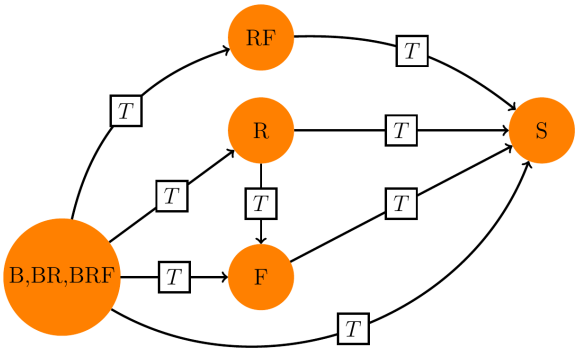

Figure 1: Pig Production Chain displayed as a directed network. The node names represent B: breeding farm, R: nursery farm, F: fattening farm, S: slaughterhouse and mixed holdings, i.e. BR: breeding/nursery farm, RF: nursery/fattening farm and BRF: breeding/nursery/fattening farm. The edges reflect livestock trade between holdings where a T indicates that a trader could be involved in the pig movement process.

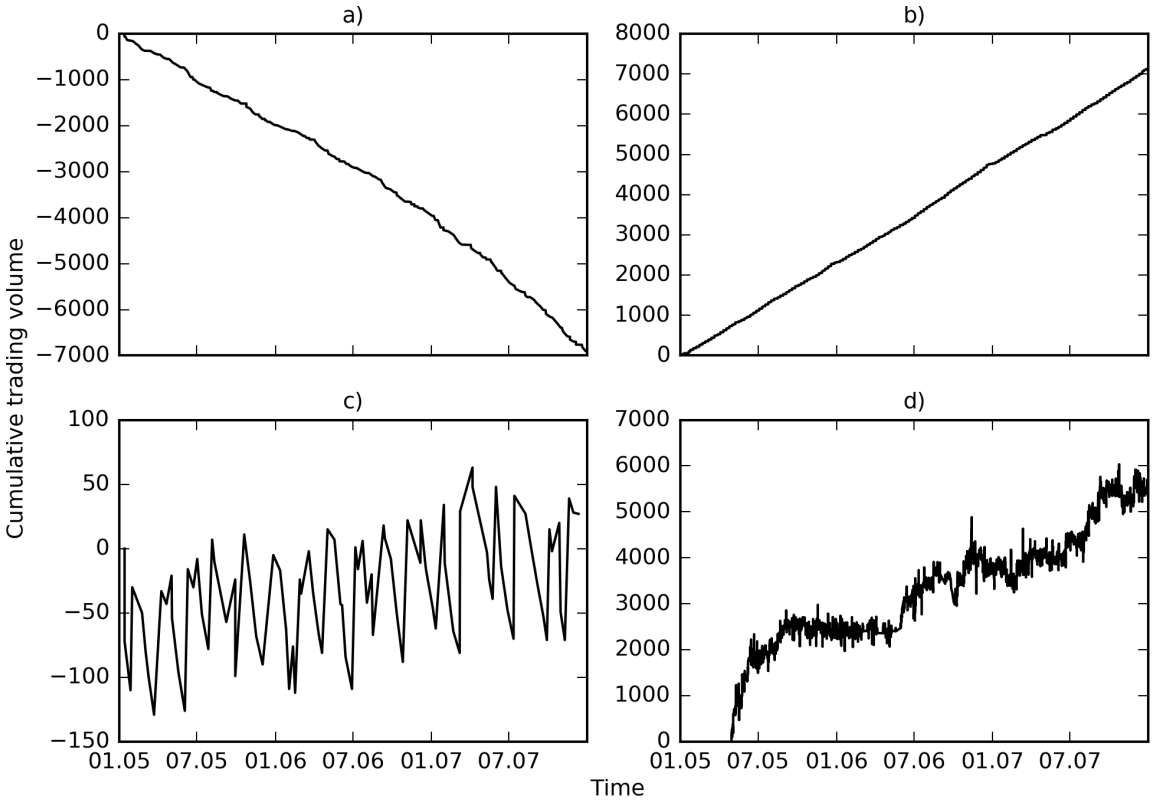

Figure 2: Cumulative number of traded pigs over time for different holding types. Four trading behavioral patterns could be identified which are characteristic for: a) breeding farm, b) slaughterhouse, c) fattening farm and d) trader. In each plot the initial number of pigs on January 1, 2005 is set to 0.

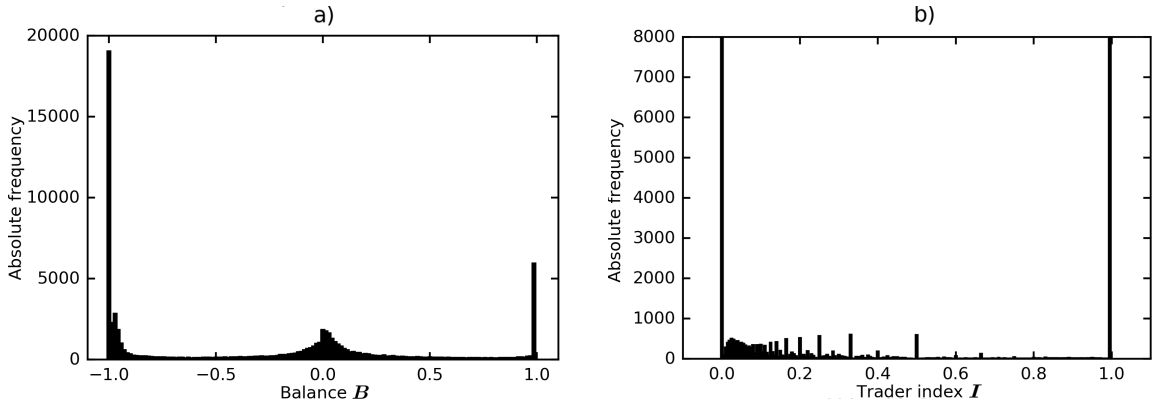

Figure 3: a) histogram of the Balance $\boldsymbol{B}$. Three modes can clearly be identified at -1, 0 and 1 which represent breeding farms, fattening farms and traders, and slaughterhouses respectively. b) histogram of the Trader Index $\boldsymbol{I}$. It shows, that the agents can be divided into two classes, namely traders and some breeding farms with a high Trader Index and non-traders with a low Trader Index.

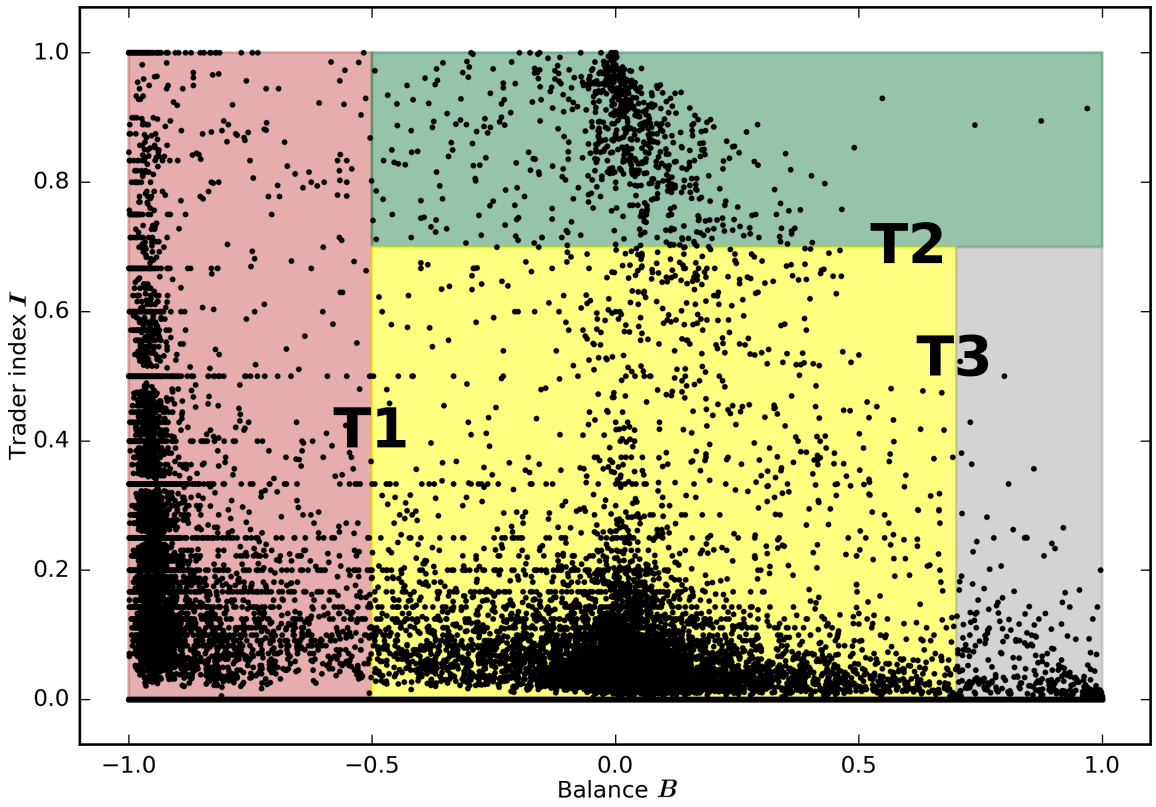

Figure 4: Classification of the holdings with thresholds $T_1 = -0.5, T_2 = 0.7$ and $T_3 = 0.7$. The colors indicate breeding farms (red), fattening farms (yellow), slaughterhouses (grey) and traders (green).

	Day t		Day $t+1$	Day $t+2$
Purchase Index	i	$i+1$	$i+2$	N/A
Purchase Volume	2	6	6	N/A
Sales Volume	3		5	0

Assignments	$f(i)$	$f(i+1)$	$f(i+2)$	
	0	0	0	} Valid
	1	0	0	
	0	1	0	
	1	1	0	Best valid
	...	...	1	Not valid

Table 1: An excerpt of an exemplary farms trading pattern and an illustration of validity. An assignment can only be valid if it assigns zero to the purchase with index $i+2$. This is because its 6 pigs cannot be sold on the same or on the very next day while only 5 pigs are available to sell according to the data. The green row marks the assignment resulting in the highest number of valid prompt purchases, which can be reached by just looking at this extract.

	Classified algorithmically				Classified by sight	
Type	abs No w.r.t 70,735	% w.r.t 70,735	abs No w.r.t 1,826	% w.r.t 1,826	abs No w.r.t 1,826	% w.r.t 1,826
$\mathcal{B}$	32,322	45.7	865	47.4	879	48.1
$\mathcal{T}$	714	1.0	15	0.8	12	0.7
$\mathcal{F}$	29,343	41.5	700	38.3	688	37.7
$\mathcal{S}$	8,356	11.8	246	13.5	247	13.5

Table 2: Comparison of the results of the algorithmic classification with the classification by sight. Regarding the sample of only 1,826 holdings, both classification methods gave approximately the same proportional sizes for each holding type. Despite the false data, the proportions of the groups regarding the whole dataset and regarding the selected sample of holdings vary only slightly.

Supplementary material to:
Automatic classification of farms and traders in the pig production chain

Lisa Köppel[a,*], Tobias Siems[b], Mareike Fischer[b], Hartmut H. K. Lentz[c]

[a] *CHICAS, Faculty of Health and Medicine, Lancaster University, Lancaster LA1 4YG, United Kingdom*
[b] *Department of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Straße. 47, 17489 Greifswald, Germany*
[c] *Friedrich-Loeffler-Institut, Institute of Epidemiology, Südufer 10, 17493 Greifswald - Insel Riems, Germany*

Introduction

This is the supplementary material to the paper "Automatic classification of farms and traders in the pig production chain" published in the Journal "Preventive Veterinary Medicine". The supplement also contains an R package that computes the Balances and Trader Indices of the holdings in given trade data.

A: Mathematical concept of the trader index

We discuss the highest trader index under all valid assignments $\boldsymbol{I}$ as introduced in Chapter 2. We fix an arbitrary premise with ℓ purchases. Let p_i be the number of pigs of purchase i and $P_t \subseteq \{1, ..., \ell\}$ the set of purchases on day t. We further define the number of pigs sold on day t as s_t.

Consider assignments $f : \{1, ..., \ell\} \to \{0, 1\}$, where $f(i) = 1$ respectively $f(i) = 0$ states that i is respectively is not sold promptly. Our first aim is to develop a mathematical concept of validity for assignments. Roughly speaking, an assignment is valid if we do not sell more pigs than we have purchased. Since an assignment f makes statements only for prompt purchases, an i with $f(i) = 1$ means that the data allows to sell p_i pigs within two days. Otherwise f cannot be valid. Therefore, $f(i)p_i - s_t \leq s_{t+1}$ must apply if the purchase i occurred on day t. Furthermore, this must also hold for all prompt purchases on day t, thus $\sum_{i \in P_t} f(i)p_i - s_t \leq s_{t+1}$ must apply as well. Moreover, we also have to consider the carry from day $t-1$, i.e. the promptly purchased pigs from day $t-1$ that are not sold on day $t-1$ and hence must be sold on day t.

Given an assignment f, we define the *carry* $\mathcal{C}_f(t)$ from day t recursively as follows

$$\mathcal{C}_f(t) := \max\left\{0\ ,\ \sum_{i \in P_t} f(i) \cdot p_i + \mathcal{C}_f(t-1) - s_t\right\},\ t = 1, ..., n$$

where $\mathcal{C}_f(0) = 0$. $\mathcal{C}_f(t)$ expresses the number of pigs that are not sold on day t and thus, have to be carried over to day $t+1$ and sold on that day . This definition exposes the recursive nature of validity and we are now able to express the precise meaning of it.

Definition 1: An assignment f is called *valid* if $\mathcal{C}_f(t) \leq s_{t+1}$ for all $t \in \{1, ..., n\}$, where $s_{n+1} = 0$.

We now want to find an efficient algorithm to compute the highest trader index under all valid assignments f. This is done through a dynamic programming algorithm by defining dynamic variables, that can be seen as small subproblems which are solved recursively. Having solved all subproblems, a solution to the main problem can be derived easily (Viterbi, 1967; Bellman, 1957).

Given a day t and a sales volume s we define $\mathcal{G}(t, s) := \max_{A \subseteq P_t}\left\{\#A \mid \sum_{i \in A} p_t \leq s\right\}$. $\mathcal{G}(t, s)$ represents the largest number of purchases that could be sold on day t if there was a sales volume of s available. Furthermore, we define dynamic variables r_{tj} for $t = 1, ..., n$ and $j = 0, ..., s_{t+1}$ through

$$r_{tj} := \max_{i \in \{0, ..., s_t\}}\left\{r_{t-1i} + \mathcal{G}(t, s_t - i + j)\right\} \tag{1}$$

*Corresponding author
Email addresses: koeppel.lisa@gmail.com (Lisa Köppel), Tobias.Siems@yahoo.com (Tobias Siems), email@mareikefischer.de (Mareike Fischer), Hartmut.Lentz@fli.de (Hartmut H. K. Lentz)

where $r_{0j} = 0$ for all j.

These dynamic variables represent our subproblems. Solving all of them recursively leads to $\boldsymbol{I}$. This is justified by the following theorem:

Theorem 1: $\boldsymbol{I} = \frac{1}{\ell} r_{n0}$

We will proof this theorem in Section A.1 following a small example:

Example 1:

The following table represents the trade of an artificial agent:

	Day 1		Day 2	Day 3	Day 4		
p_i	1	6	4	8	1	2	3
s_t	5		3	7	4		

With this we can compute the r_{tj}'s according to Equation (1), which is illustrated in the following table:

t \ j	0	1	2	3	4	5	6	7
1	1	1	2	2				
2	2	2	2	3	3	3	3	3
3	3	3	3	3	4			
4	5							

This states that $\boldsymbol{I}$ equals $\frac{5}{7} \approx 0.714$, which means that up to 71% of the purchases can be considered as prompt without violating validity. Since this is a high number, we should regard this agent as a trader.

A.1 Proof of Theorem 1

First of all, we need two definitions that are inspired by the r_{tj}.

Definition 2: A valid assignment $f : \{1, ..., \ell\} \to \{0, 1\}^l$ is called (t, j)-*valid* if

$$\mathcal{C}_f(t) \leq j \text{ and } k > t,\ i \in P_k \text{ implies } f(i) = 0$$

Furthermore, a (t, j)-valid assignment f is called *maximal* (t, j)-*valid* if $\boldsymbol{I}^h \leq \boldsymbol{I}^f$ for all (t, j)-valid assignments h.

(t, j)-valid assignments never declare purchases on days past t as prompt, and they are only allowed to have a carry of at most j pigs that must be sold on day $t + 1$. An $(n, 0)$-valid assignment is valid and the trader index of a maximal $(n, 0)$-valid assignment equals $\boldsymbol{I}$. These definitions are closely connected to the r_{tj}'s as shown in the following lemma.

Lemma 1: r_{tj} is the trader index of any maximal (t, j)-valid assignment.

Proof of Lemma 1. Since all maximal (t, j)-valid assignments have the same trader index, it is sufficient to proof the claim for one maximal (t, j)-valid assignment. We perform a proof by induction with respect to $t = 1, ..., n$. Therefore, let r_{t-1j} be the trader index of a maximal $(t - 1, j)$-valid assignment and let f be a maximal (t, j)-valid assignment. We notice that $\boldsymbol{I}^f = \sum_{i \in P_t} f(i) + \sum_{i \notin P_t} f(i)$. Since f is maximal, $\sum_{i \in P_t} f(i) = \mathcal{G}(t, s_t - \mathcal{C}_f(t - 1) + j)$ must apply. Furthermore, since f is maximal, $r_{t-1,\mathcal{C}_f(t-1)} = \sum_{i \notin P_t} f(i)$ applies by the induction hypothesis. Finally, we have to show that $r_{t-1i} + \mathcal{G}(t, s_t - i + j) \leq r_{t-1,\mathcal{C}_f(t-1)} + \mathcal{G}(t, s_t - \mathcal{C}_f(t - 1) + j)$ for all $i = 0, ..., s_t$. However, this is also justified by the fact that f is maximal (t, j)-valid.

□

With this, we also have a proof of Theorem 1.

Algorithm 1

```
1:  function 𝒢(t, s)                                          ▷ Compute 𝒢(t, s)
2:      list L={p_t | t ∈ P_t}                                ▷ L holds a list of volumes purchased on day t
3:      sortAscending(L)
4:      d = 0
5:      for i = 1, ...,length(L) do                           ▷ Iteration over the length of L
6:          d = d+L[i]
7:          if d > s then                                     ▷ Find the largest number of purchases with a volume of at most s
8:              return i − 1
9:          end if
10:     end for
11:     return length(L)
12: end function
13:
14: function COMPUTE I( )
15:     for j = 0, ..., s_1 do
16:         r_0j = 0                                          ▷ Declare r_0j to be zero for all j
17:     end for
18:     for t = 1, ..., n do                                  ▷ Iteration over days
19:         for j = 0, ..., s_{t+1} do                        ▷ Iteration over purchase volume from day t + 1
20:             r_1j = 0
21:             for i = 0, ..., s_t do                        ▷ Iteration over purchase volume from day t
22:                 d=r_0i+𝒢(t, s_t − i + j)
23:                 if d > r_tj then                          ▷ Find maximum according to Equation (1)
24:                     r_1j = d
25:                 end if
26:             end for
27:         end for
28:         for j = 0, ..., s_{t+1} do
29:             r_0j = r_1j                                   ▷ Swap old values
30:         end for
31:     end for
32:     return r_10                                           ▷ Return I
33: end function
```

B: An algorithm to compute the highest trader index

Now we develop an algorithm that computes r_{tj} for $t = 1, ..., n$ and $j = 0, ..., s_t$ and derive its time and space complexity. Having computed all r_{tj}'s we receive $\boldsymbol{I}$ from r_{n0} as stated in Theorem 1. Algorithm 1 shows the pseudocode of our algorithm.

$\mathcal{G}(t, s)$ has a time complexity of $\mathcal{O}(\#P_t \cdot \log(\#P_t))$, which comes from the sorting in line 4. The space complexity is $\mathcal{O}(\#P_t)$. The function compute$\boldsymbol{I}$ has a time complexity of $\mathcal{O}(n \cdot \max\{s_t\}^2 \cdot \max\{\#P_t\} \cdot \log(\max\{\#P_t\}))$, which comes from the nested loops in 19, 20 and 22 and the calculation of $\mathcal{G}(s, t)$ in 23. The space complexity is $\mathcal{O}(\max\{s_t\})$.

If we consider the trade as limited regarding the maximal sales volume and the maximal number of purchases per day, then all the above complexities are at most linear in n. This means that our algorithm scales well with very large datasets and should generally be applicable with moderate hardware in use.

C: Choosing the thresholds

In the paper, we introduced the Balance and the Trader Index to identify different trading patterns. By means of these two indices we now turn our focus to the automatic classification of all agents given in our dataset. In particular, we want to find suitable thresholds for the indices in order to define four groups which are illustrated in a scatter plot of the two indices in Figure 1.

In order to choose suitable values for the thresholds we proceed in the following way: For threshold T_1 we first consider the number of holdings in the class of breeding farms when shifting the threshold from -1 to 1 as displayed by the green line in Figure 2 a). It becomes apparent that we can observe an inflection point at around -0.5. This describes a minimum of the growth rate of the number of breeding farms and can be explained as follows: To the left of -0.5, the contribution to the cumulative number of breeding farms decreases due to a declining number of fattening farms, while to the right of -0.5, the contribution increases due to a rising number of fattening farms, traders and later slaughterhouses. Furthermore, since breeding farms sell a small amount of replacement gilts to other breeding farms (within the same class), we analogously considered the ratio of transactions between breeding farms with regard to all transactions in the dataset while shifting the threshold (blue line). Again, we can clearly identify an increase of the curve at around -0.5. Hence, it was reasonable to choose the inflection point to set the threshold $T_1 = -0.5$.

For setting threshold T_2, we consider the fraction of slaughterhouse sales with respect to all transactions in the dataset (Figure 2 b). Since a slaughterhouse is located at the end of the pig production chain, the

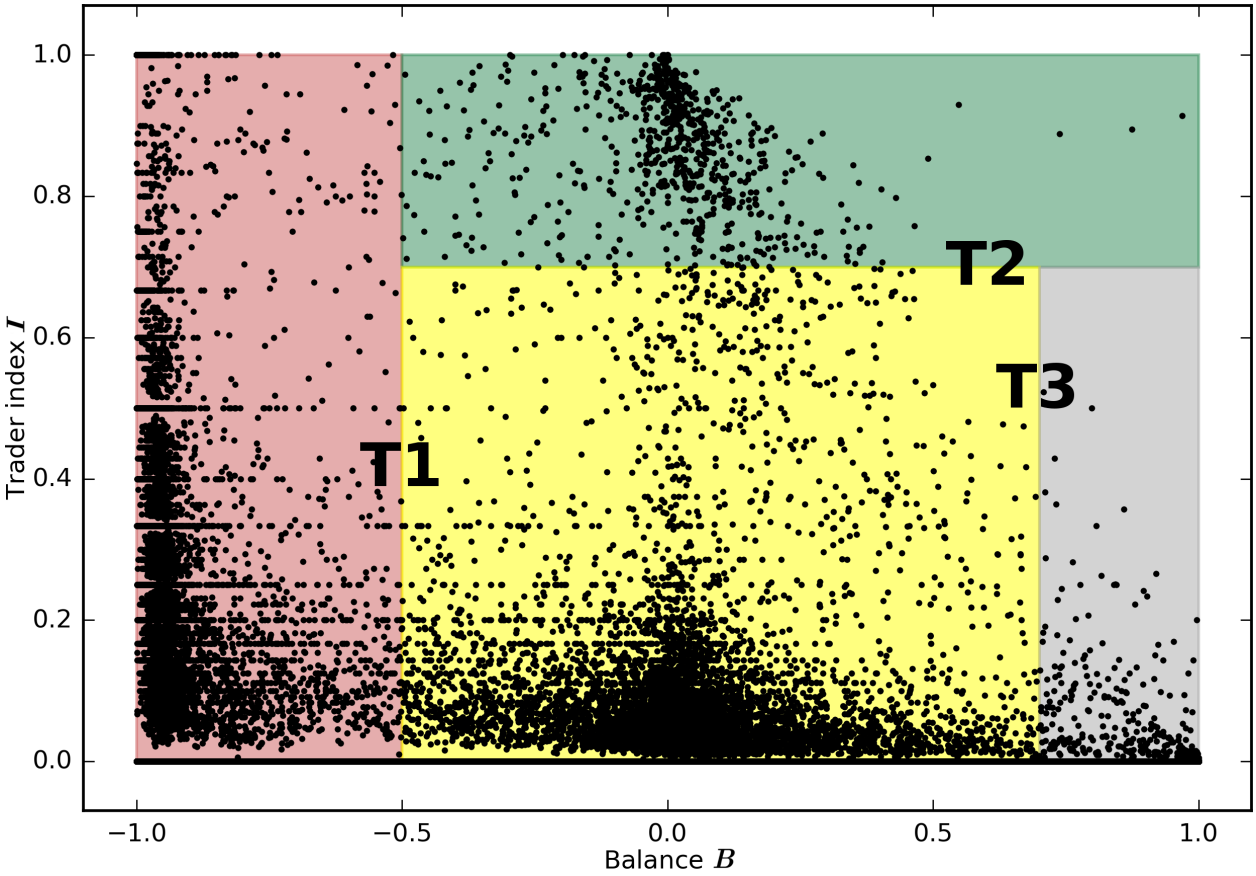

Figure 1: Classification of the holdings with thresholds $T_1 = -0.5, T_2 = 0.7$ and $T_3 = 0.7$. The colors indicate breeding farms (red), fattening farms (yellow), slaughterhouses (grey) and traders (green)

value for its sales fraction has to be chosen close to zero. By setting $T_2 = 0.7$, we achieve a fraction of slaughterhouse sales of only 0.68% as desired. It should be noted that we refrained from solely defining holdings that only purchased pigs as slaughterhouses. Instead, our goal is to classify all farms with a similar behavior, here heavy on sales, into one group.

The threshold T_3 is to be set in such a way that it isolates the point cloud at the top in the center of the scatter plot in Figure 1. In order to separate traders especially from fattening farms, we drew a histogram (Figure 3) of only the green and the yellow area in the scatter plot. It would be sensible to set $T_3 = 0.6$ as the number of holdings seems to increase thereafter. However, as we denote traders to have a high number of prompt purchases, we set the threshold a little higher, namely at $T_3 = 0.7$.

Further justification of the specific choice of T_3

We want to examine our threshold T_3, which was chosen to be slightly more vague than the other two thresholds. For this reason, we plotted an ROC plot in Figure 4 with the sample of size 1,000 which was mentioned at the end of the results section in the paper. It illustrates the true positive rate against the false positive rate at shifting threshold settings. It displays that when shifting the threshold T_3 from 0 to 1, we first observe sensitivity values of 1.0, as expected, with a decreasing false positive rate. Exactly at 0.7 (marked by the red cross), we reach a false positive rate of 0 with following significant drops in sensitivity. Hence, this plot shows that setting $T_3 = 0.7$ is low enough to include as many traders as possible while impairing the true positive rate only very slightly.

References

Bellman, R. (1957). *Dynamic Programming*. (1st ed.). Princeton, NJ, USA: Princeton University Press.

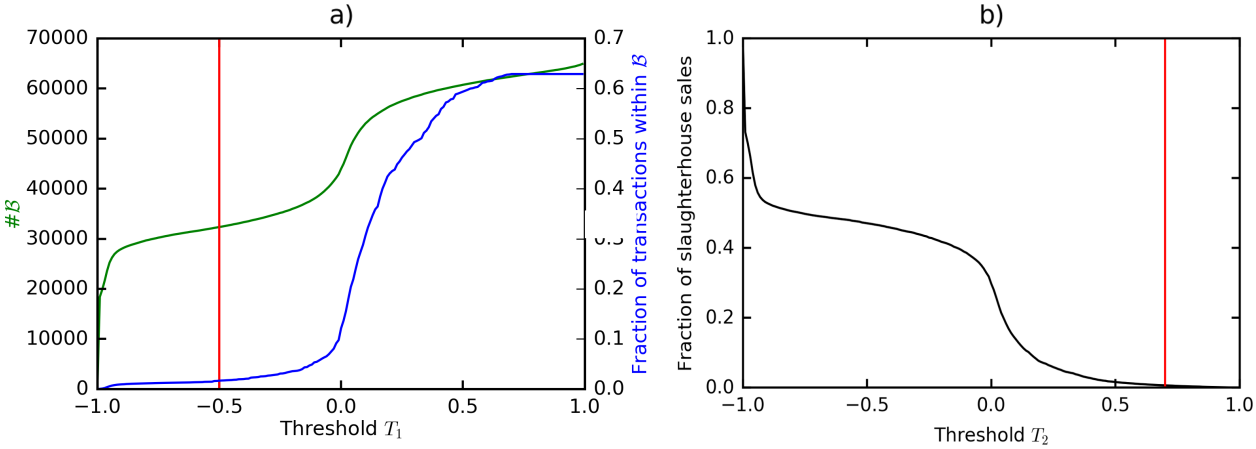

Figure 2: Determining thresholds T_1 (a) and T_2 (b). For T_1 we displayed the number of breeding farms (green) and the fraction of transactions within the class $\mathcal{B}$ with respect to all transactions (blue) while shifting the threshold. The vertical red line displays the chosen value for T_1 at the estimated inflection point. Threshold T_2 is set in such a way that the fraction of slaughterhouse sales with respect to all transactions is almost equal to 0 while shifting the threshold.

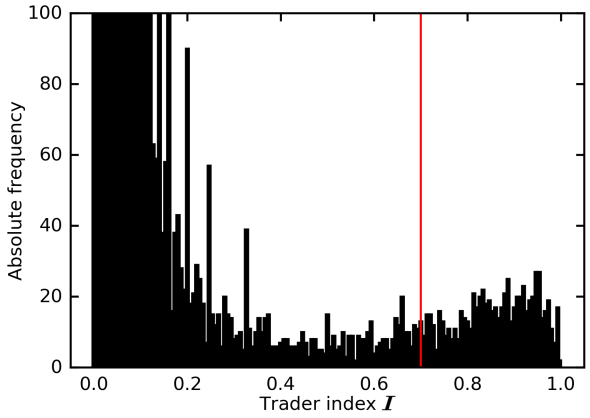

Figure 3: Histogram of the Trader Index of holdings with $-0.5 < \boldsymbol{B} \leq 0.7$. The vertical red line displays the chosen value for T_3, which separates the class of fattening farms from the class of traders.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*.

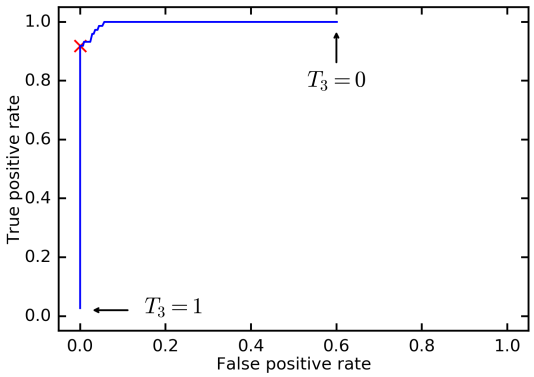

Figure 4: ROC Plot of trader classification with shifting threshold T_3. The red cross marks the point of the chosen threshold $T_3 = 0.7$ with a false positive rate of 0 and following significant drops in sensitivity. This affirms the choice of its value because as many traders as possible were included while sacrificing the true positive rate only very slightly.