Social Networks 54 (2018) 73–81

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Effects of data quality in an animal trade network and their impact on centrality parameters

Kathrin Büttner*, Jennifer Salau, Joachim Krieter

Institute of Animal Breeding and Husbandry, Christian-Albrechts-University, Olshausenstr. 40, D-24098 Kiel, Germany

ARTICLE INFO

Article history:

Keywords:
Network analysis
Trade network
Pig
Boundary specification problem
Targeted removal scenarios

ABSTRACT

Dealing with the analysis of animal trade networks always faces the challenge of imperfect data sets mainly due to country borders or different producer communities. In the present study, the network robustness, i.e. the point at which false positive nodes or edges may influence the network structure and the results of the centrality parameters, were analysed for a pork supply chain of a producer community in Northern Germany. The analysis of animal trade networks mainly focusses on disease transmission and the development and implementation of targeted prevention and intervention strategies based on centrality parameters. Here, the inclusion criteria may impact the prediction of disease transmission as well as the outcome of the applied control measures. Thus, four different removal scenarios all based on the boundary specification problem (removal of arcs according to their frequency of appearance, removal of nodes according to their general frequency of appearance and according to their frequency of appearance as supplier or purchaser) were established to analyse the network robustness. In order to evaluate the changes in the rank order of the nodes a Spearman Rank Correlation Coefficient (r_s) was calculated between the original network and each removal step. The removal of nodes according to their frequency of appearance showed the most robust results. The values of r_s stayed above the threshold of 0.70 for at least a fraction of 80% removed arcs. For the other removal scenarios the centrality parameters under investigation showed various robust results concerning the ranking of the nodes.

Therefore, the exclusion of farms that trade infrequently in the network would not be associated with significant change in network structure and centrality parameters. For targeted disease prevention and intervention strategies based on centrality parameters, it is of great relevance to be able to evaluate the influence of inclusion criteria on the network structure and thus on the speed and the extent of possible disease transmission.

1. Introduction

Network theory has become a valuable framework in many different research areas. For instance, in the social sciences, e.g. contacts of individuals (Büttner et al., 2015b,c ; Kasper and Voelkl, 2009; Krause et al., 2007; Lewis et al., 2008; Makagon et al., 2012), in epidemiological studies, e.g. disease transmission over trade networks (Bigras-Poulin et al., 2007; Büttner et al., 2013a, 2016 ; Konschake et al., 2013; Lentz et al., 2016; Nöremark et al., 2011; Rautureau et al., 2012) or in the analysis of technological networks, e.g. the World Wide Web or the internet (Albert et al., 1999; Barabási et al., 2000; Cohen et al., 2000), to name but a few.

However, dealing with network analysis always faces the challenges associated with imperfect data, e.g. some nodes or edges are missing or information about some of their attributes are missing, and issues related to inclusion criteria for nodes or edges, e.g. nodes or edges which are not actual members of the network under investigation are included (false-positives). According to Kossinets (2006), much of this imperfection arises from the following sources: the so-called boundary specification problem (Laumann et al., 1983) as well as inaccuracy or non-response to data recording, which is especially valid for surveys (Brewer and Webster, 1999; Butts, 2003; Robins et al., 2004; Stork and Richards, 1992).

For the analysis of animal trade networks primarily with focus on disease transmission and the development of prevention and control strategies, data quality and inclusion criteria used may influence the network structure and therefore the prediction of disease transmission as well as the outcome of the applied control measures. Only if the ranking of the farms based on the centrality parameters remained stable, the results of the network analysis can be used as reliable indicators for targeted removal strategies

* Corresponding author.
E-mail address: kbuettner@tierzucht.uni-kiel.de (K. Büttner).

https://doi.org/10.1016/j.socnet.2018.01.001
0378-8733/

in the case of an epidemic. This means that the ranking of the centrality parameters must not change easily if only a small amount of elements are removed from or added to the network to ensure reliable results. In the case of pig trade networks, country borders or specific producer communities are often used as the inclusion criteria when building these networks, which can be associated with false-positives. In addition, data quality and inclusion criteria may be affected by the fact that additional information such as farm type or size are recorded only in specific marketing programs in which only a small amount of farms participate.

The boundary specification problem (Laumann et al., 1983) deals with the issue of specifying system boundaries. The outcome of the network analysis depends on the nodes and edges included in the whole system, meaning special care must be given to specify the rules of inclusion of network elements (both nodes and edges). In the case of pig trade networks, the question which has to be answered is which farms or trade contacts should be qualified as legitimate members of the producer community under investigation? It can be assumed that network elements (i.e. nodes and edges) which appear less frequently in the network in the time period under analysis do not really belong to the studied producer community but sold their animals via this specific producer community only exceptionally. Therefore, these can be considered false positive nodes resulting from the supply and demand or different payment conditions within the pork supply chain. According to Barnes (1979), the centrality parameters may be underestimated if a too restrictive boundary is chosen. Additionally, in accordance to Wang et al. (2012), false positive nodes or edges, i.e. elements that are erroneously present in the network, can have a great impact on the outcome of the network analysis. Although one often encounters different data qualities, little research has been carried out to evaluate the effect of false positive nodes or edges on the network-level or node-level parameters (Kossinets, 2006).

Therefore, the aim of this study was to evaluate the network robustness under different network boundary definitions, i.e. to analyse how the rules of inclusion for nodes or edges in a network may influence the centrality parameters (e.g. in-degree and out-degree centrality, betweenness centrality, ingoing and outgoing closeness centrality). To carry out a sensitivity analysis, different removal scenarios representing various rules of inclusion and thus different exclusion criteria were established. For each removal step the development of the centrality parameters was recorded and then compared with each other. Especially for disease control strategies based on centrality parameters, it is of great relevance to understand the influence of the rules of inclusion, or how the boundary of the network is defined on the network structure and thus on the impact on the speed and the extent of possible disease transmission.

2. Materials and methods

2.1. Data basis and network construction

In an observation period from 1st January 2013 to 31st December 2014 pig trade data from a producer community in Northern Germany were recorded. The producer community organizes the marketing of live pigs for their members and in this context also registers all trade contacts between its members, i.e. the transportation of live animals between a supplier and a purchaser on a given day. Here, the suppliers and purchasers were the nodes of the network which were connected by the trade contacts, i.e. the edges of the network. Each trade contact had one specific supplier and one specific purchaser, i.e. each edge has a certain direction, hereafter referred to as arcs. Furthermore, a farm was categorized as supplier if it delivered animals to another farm and it was categorized as purchaser if it got animals from another farm. Due to this categorization there were also farms which could be both supplier and purchaser.

In addition to the complete network which covered the whole observation period, yearly networks were constructed. For the yearly networks, the description of the network topology as well as the results of the sensitivity analysis can be found in the Supplementary material.

Out of all the daily records a static aggregated network was constructed. Therefore, repeated trade contacts between the same farms during the observation period were aggregated to a single one.

2.2. Boundary specification problem

All farms and trade contacts which were listed in the recordings were firstly allocated to the producer community in Northern Germany used for this study. According to Borgatti and Halgin (2011) it is important to realize that due to these settings the network itself and simultaneously its boundaries were predefined. The question is, however, if these boundaries were properly chosen to correctly describe the producer community under investigation or if the data set contains so-called false positive nodes or arcs. This is exactly the issue addressed by the boundary specification problem (Laumann et al., 1983). It deals with the question which set of units should be included in the network, i.e. the choice of the right exclusion or inclusion criteria of network elements. According to Marsden (1990), this is comparable to the general problem of defining the population to which research results are to be generalized.

In the case of the pig trade network of the present producer community the question that has to be answered is: Do the farms or trade contacts really belong to this producer community, or were they recorded in the data set for some other reasons? For this purpose it was assumed that nodes or arcs which occurred less frequently in the analysed time periods were not legitimate members of the producer community under investigation. One explanation for their low occurrences in the data set might be that these farms sold their animals via this specific producer community only exceptionally. Thus, they can be seen as artefacts which may result because of supply and demand or different payment conditions within the pork supply chain.

If the inclusion or exclusion criteria and therefore also the system boundaries are chosen too restrictive, the network structure may change which also influences the results of the centrality parameters (Barnes, 1979). In addition, false positive nodes or edges, i.e. elements that are erroneously present in the network, have a great impact on the outcome of the network analysis (Wang et al., 2012). In the case of the producer community under investigation, false positive nodes or edges can be considered as these artefacts.

2.3. Definition and implementation of removal scenarios

2.3.1. Definition of removal scenarios

Based on the above described considerations four removal scenarios were established to investigate the influence of false positive nodes or arcs on the outcome of the network analysis: removal scenario 1 (removal of trade contacts according to their frequency of occurrence in the data set), removal scenario 2 (removal of farms according to their frequency of occurrence in the data set), removal scenario 3 (removal of suppliers according to their frequency of occurrence in the data set) and removal scenario 4 (removal of purchasers according to their occurrence in the data set).

The distinction between the removal of nodes and arcs was made because of the hypothesis that the removal of nodes may have