Network-based Source Detection: From Infectious Disease Spreading to Train Delay Propagation

Juliane Manitz\textsuperscript{1}, Jonas Harbering\textsuperscript{2}, Marie Schmidt\textsuperscript{2}, Thomas Kneib\textsuperscript{1}, Anita Schöbel\textsuperscript{2}

\textsuperscript{1} Department of Statistics and Econometrics, Georg-August University Göttingen, Göttingen, Germany
\textsuperscript{2} Institute for Numerical and Applied Mathematics, Georg-August University Göttingen, Göttingen, Germany

E-mail for correspondence: jmanitz@uni-goettingen.de

Abstract: The correct identification of the source of a propagation process is crucial for research questions in a wide range of research fields: to determine the epicenter of infectious disease outbreaks, the onset of blackouts in power grids, the root of computer virus attacks in the Internet, the origin of misinformation in social networks, or the starting point of the invasion of non-endemic species in ecology. Here, we consider source determination of train delays in railway systems, which mimic many-faceted diffusion patterns. Delays can never be entirely avoided, but their impact has to be kept to a strict minimum. We enhance a fast and efficient approach for the source identification of propagation processes on networks, which is structurally quite general and only requires a minimum data basis. In extensive simulation studies, we investigate the performance in dependency of time and network node centrality. We examine the robustness of the approach by the application of different delay management strategies, which mimic various propagation mechanisms. Finally, we test for performance improvement due to the integration of additional knowledge in the network definition. Altogether, the source detection framework turns out to be robust for diverse spatio-temporally evolving processes, which promises the general applicability in many research fields.

Keywords: Source Detection; Complex Network; Public Transportation.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
1 Introduction

Many spreading phenomena, e.g., the transmission of diseases and the propagation of delays in railway networks can be modeled as processes on networks. The aim of source detection is to find the starting point of such a propagation process from data about the observed event counts at the network nodes. With the knowledge of the origin of a propagation process, one is able to truly combat further spreading. Additionally, the origin is the basis for the prediction of the propagation process. Therefore, source detection plays a crucial role in the problem assessment of spreading phenomena.

Thereby, modern propagation patterns are highly complex and irregular. They can be described best by processes on complex networks. Therefore, we enhance the network-based approach for source detection by Manitz et al. (2014), which has been originally developed to reconstruct the epicenter of food-borne disease outbreaks. As a many-faceted application, we chose delay propagation in railway networks. Based on a well-defined network, the application has the advantage that good models for delay propagation exist. Thus, the spreading of delays can be easily simulated and various complex diffusion patterns can be mimicked. Hence, delay propagation on railway networks is a good candidate example to test whether the network-based approach can be applied for source detection problems other than the spreading of food-borne diseases.

2 Methods and Data

2.1 Network-Based Source Detection

The approach requires an underlying network, which can be specified as a collection of nodes $k = 1, \ldots, K$, which are connected by direct links between them. When modeling food-borne infectious diseases, the underlying network represents the transportation routes of contaminated food. Here, the network is defined by a public transportation system, where nodes represent railway stations. Two nodes are connected by a link (also called edge) if there is a track between the corresponding stations, which is used by a scheduled train.

An appropriate distance $d(k, l)$ is defined for a specific path $\gamma_{kl}$ between all pairs of nodes $k, l = 1, \ldots, K$ of the network. This will be specified in the next section. Furthermore, we assume a time-dependent stochastic process $\{X_k(t)\}$ on the network nodes characterizing a propagation mechanism in a time range $t = 1, \ldots, T$. Corresponding observations $x_k(t)$ in each node $k$ are conducted at different time points $t = 1, \ldots, T$ to find sequential pictures of the distribution pattern.

The basic assumption is that propagation phenomena are spreading in a circular pattern from the correct origin $k_0$. The focal idea of source recon-
struction is testing different source candidates and examine the concentricity of the observed pattern on a minimum shortest-path tree with the candidate $k_0$ as the root. Thus, given an appropriate distance definition $d$, the source can be reconstructed as the median of the observed pattern at time $t$, which is obtained by minimizing the expected distance $\mu_X(d; k_0, t)$ from the origin $k_0$ to all other network nodes, i.e.

$$\hat{k}_0(t) \in \arg \min_{k_0 \in \mathcal{K}_0} \mu_X(d; k_0, t), \quad (1)$$

where $\hat{k}_0(t)$ is from the set of nodes $k_0 \in \mathcal{K}_0$ for which the expected distance attains the smallest value, i.e. $\mu_X(d; \hat{k}_0, t) = \min_{k_0 \in \mathcal{K}_0} \mu_X(d; k_0, t)$.

Thereby, the expected distance $\mu_X(d; k_0, t)$ can be estimated by the average distance $d(k, k_0)$ from source $k_0$ to all destination nodes $k$ weighted by the observed number of delays $x_k(t)$ in node $k$ until time $t$. Thus,

$$\hat{\mu}_X(d; k_0, t) = \frac{1}{N_x(t)} \sum_{k=1}^{K} x_k(t) \cdot d(k, k_0), \quad (2)$$

where $N_x(t) = \sum_k x_k(t)$ is the total number of delays in the network at time $t$.

For the transformation of the irregular diffusion pattern into a typical concentric spreading circle, the replacement of the classic geographic distance by a network-based effective distance is necessary (see Brockmann and Helbing, 2013; Manitz et al., 2014).

### 2.2 Characteristics of the Railway Network

The public transportation network consists of $K = 319$ nodes connected by 446 links, which results in a very low link density of 0.009. Only about 1% of all possible links in a fully connected network are present (see Figure 1). The average link number to other stations is 2.8 and similar to other PTNs (e.g., von Ferber et al., 2009). The majority of the stations are stops on a line (median is 2). The degree distribution is left-skewed, so that there are a few stations of high importance with a large number of links in various directions.

### 2.3 Train Delay Simulation

Based on a public transportation network (see Section 2.2) we compute a line concept and a timetable. We generate a set of initial delays, which model exterior influences such as weather conditions, strikes or construction work. Those initial delays are then propagated, because of dependencies between the trains due to passenger transfers or track occupation of subsequent trains. The decision which passenger transfers can be hold and
the sequence of trains running along a track are made according to a prescribed delay management strategy. They allow to remove transfers from the delayed trains and to switch train orders in order to decrease the impact of delays. Using the LinTim software package (Goerigk et al., 2014) for executing delay management strategies we are able to generate diverse propagation mechanisms to mimic various interesting spreading patterns.

3 Results and Conclusions

The simulation results confirm the applicability of the source detection approach, that suggest that train delay spreading has similar underlying propagation mechanism as the transmission of infectious diseases. We observe decreasing source detection performance over time, while the influence of node centrality is moderate, if regular networks are considered. It can be also shown that our approach for source detection is extremely robust in regard to different propagation mechanisms. Furthermore, the incorporation of additional knowledge in the network definition improves the source detection performance. However, the unweighted network performs only a little worse, so that the approach can be recommended even without knowledge for link weighting. The simulation results illustrate the applicability of the method not only in the area of infectious diseases but also in the area of train delays.
Acknowledgments: This work was supported by the German Research Foundation, Research Training Group 1644 'Scaling Problems in Statistics' and the Simulation Science Center Clausthal/Göttingen.

References


