View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Erasmus University Digital Repository

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Applied Statistics
Series C

Appl. Statist. (2017)
66, *Part* 3, *pp.* 521–536

Source estimation for propagation processes on complex networks with an application to delays in public transportation systems

Juliane Manitz,

Georg-August-Universität Göttingen, Germany, and Boston University, USA

Jonas Harbering,

Georg-August-Universität Göttingen, Germany

Marie Schmidt

Erasmus University Rotterdam, The Netherlands

and Thomas Kneib and Anita Schöbel

Georg-August-Universität Göttingen, Germany

[Received September 2015. Final revision June 2016]

Summary. The correct identification of the source of a propagation process is crucial in many research fields. As a specific application, we consider source estimation of delays in public transportation networks. We propose two approaches: an effective distance median and a backtracking method. The former is based on a structurally generic effective distance-based approach for the identification of infectious disease origins, and the latter is specifically designed for delay propagation. We examine the performance of both methods in simulation studies and in an application to the German railway system, and we compare the results with those of a centrality-based approach for source detection.

Keywords: Complex network; Delay spreading; Propagation process; Public transportation network; Source detection; Statistical network analysis

1. Introduction

Although train delays can never be entirely avoided, it is desirable to minimize their effect. Between April 2012 and March 2013, 20.4% of the German long-distance high-speed trains were delayed by more than 5 min (on average 15 min; see Plöchinger and Jaschensky (2013)). These delays are not only an inconvenience for passengers but also a significant economic burden for the system operator.

A key element in reducing delays in *public transportation networks* (PTNs) is the successful identification of the source of delay (also called the origin) from a specific delay pattern. To accomplish this goal, it is important to distinguish between source and propagated delays. On the basis of this, it can be investigated whether the cause of delay can be dissipated or avoided. Furthermore, the source is the basis for the prediction of the future propagation process on

Address for correspondence: Juliane Manitz, Department of Statistics and Econometrics, University of Göttingen, Humboldtallee 3, Göttingen 37073, Germany.
E-mail: jmanitz@uni-goettingen.de

© 2016 The Authors Journal of the Royal Statistical Society: Series C Applied Statistics 0035–9254/17/66521
Published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

complex networks. However, it is surprisingly complicated to track sources of delays in PTNs, because of the complex composition of delays. The recorded data are usually very inaccurate and ambiguous (Yabuki *et al.*, 2015). In this work, we consider examples of extensive delay spreading provided by the largest German railway operator: Deutsche Bahn. The delays affect large parts of the German railway network, of which the vast majority is caused in one particular station.

For a mathematical formalization of delay spreading, we assume a PTN with a line plan, and a predefined timetable. Exterior local influences, such as local construction sites, large regional demonstrations or a malfunction of a local system, introduce disturbances in the form of delays in the timetable. Those single-source delays are then propagated through the network. The delays are also transmitted to other trains, because of dependences between the trains due to passenger transfers or occupation of the track by subsequent trains. With respect to these dependences, different strategies can be chosen to avoid the spread of delays (*a delay management strategy*). A mathematical framework for describing the highly complex and irregular patterns of delay spreading on PTNs are stochastic processes on networks.

Some previous work on analysing patterns of the spread delays in PTNs was based on certain association rules (Yabuki *et al.*, 2015). Network theoretic analysis focuses on empirical investigation of structural properties, such as small world characteristics, of different types of PTN such as bus and tramway networks (Sienkiewicz and Holyst, 2005) and entire city systems (von Ferber *et al.*, 2009). A few approaches have been suggested to deduce the source of complex spreading patterns in other applications such as infectious disease epidemiology (Prakash *et al.*, 2012; Fioriti *et al.*, 2014; Pinto *et al.*, 2012; Comin and da Fontoura Costa, 2011), computer science (Shah and Zaman, 2010) or communication studies (Lappas *et al.*, 2010; Shah and Zaman, 2012; Adar and Adamic, 2005). For a comprehensive review see Jiang *et al.* (2014).

In this work, we present an approach towards source reconstruction of propagation processes with a single source based on Brockmann and Helbing (2013) and Manitz *et al.* (2014). This is a method which has initially been developed to reconstruct the origin of outbreaks of disease. On the basis of a single snapshot of the propagation process, a regular wavefront spreading is reconstructed by using an effective distance projection. In addition to this structurally generic approach for source estimation, we suggest a recursive backtracking algorithm which is specifically designed for propagation mechanisms of delays in PTNs. Note that we restrict our analysis to single-source problems. However, there are simple solutions to convert a multisource pattern into a number of single-source estimation problems (see, for example, Zang *et al.* (2014)). We compare our approaches against an adapted centrality-based method by Comin and da Fontoura Costa (2011). In an extensive simulation study, we investigate the performance of the proposed methods in the multifaceted example of the identification of sources of delay in PTNs. Based on a well-defined network, sophisticated models for delay propagation exist and are implemented, for example, in LinTim (Goerigk *et al.*, 2013), so that complex diffusion patterns can be mimicked. Since the spreading of delays is a dynamic phenomenon, the dependence on time is studied. As noise is unavoidable, the performance of these methods is studied with respect to different levels of noise. Further it is investigated whether the methods prove to be robust with respect to different network structures and propagation patterns. Finally, the effect of the centrality of source nodes is analysed.

The paper is structured as follows. In Section 2, we introduce the PTNs that are used in this study. In Section 3, the methods for source estimation are explained. Source estimation performance is analysed in an extensive simulation study in Section 4. In Section 5, we apply the methods to delay propagation data examples provided by the largest German railway operator: Deutsche Bahn. Finally, we conclude our findings in Section 6.

The data that are analysed in the simulation study can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Public transportation network data

In PTNs, nodes represent stops or stations. Two nodes are connected by a link if there is a direct connection by a scheduled line between the stops. Hence, the PTNs naturally consist of one connected component. In this work, we use delay data on four different transportation networks: three artificial data sets L1, L2 and L3, and a real world case L0 that was provided by Deutsche Bahn. We compare their structural characteristics with those from other PTNs that have been reported in the literature (Table 1). In comparison with the food shipping network, that was constructed for source detection during food-borne disease outbreaks in Germany (Manitz *et al*., 2014), we can conclude that PTNs generally seem to be very sparse, with relatively long path lengths and low transitivity. The network that was provided by Deutsche Bahn that is used in the application (see Section 5) consists of 1049 nodes and 3484 links. This network has low density, which is typical for PTNs. As a transregional system combining high speed and regional connections, the average link number is larger than in city transportation networks. The longest shortest path, i.e. the diameter, is quite low, meaning that the railway connections are efficient compared with other PTNs. For our simulation study, we use three different PTNs from the optimization software LinTim (Gattermann *et al*. (2016); see Section 4). All networks exhibit a relatively low diameter compared with other PTNs reported in the literature (Sienkiewicz and

Table 1. Empirical network characteristics for various public transportation networks in comparison with world city networks (ranges reported in von Ferber *et al*. (2009)), Polish bus and tram systems (Sienkiewicz and Holyst, 2005) and the food shipping network based on the gravity model constructed to identify the origin of food-borne disease outbreaks in Germany (Manitz *et al*., 2014)†

		Number of nodes	*Number of links*	*Density*	*Average degree*	*Average unit betweenness*	*Diameter*	*Transitivity*
Application								
German railway L0		1049	3484	0.0064	6.64	0.0031	11	0.17
PTNs from LinTim								
High-speed railway L1		319	446	0.0088	2.80	0.027	25	0.14
Göttingen bus L2		257	548	0.0083	4.26	0.096	35	0.12
Athens metro L3		51	52	0.0408	2.04	0.177	29	0.00
PTNs in the literature								
World city networks	(Minimum)	1494	5849	0.0009	2.18	0.0013	27	0.02
	(Maximum)	44629	52885	0.0018	3.73	0.0097	210	0.14
Polish bus and tram	(Minimum)	152	220	0.0010	2.53	—	—	0.03
	(Maximum)	2811	3978	0.0192	3.08	—	—	0.16
Food shipping network								
German gravity model		412	30646	0.1810	148.77	0.0066	5	0.64

†The *density* is the proportion of all actual links compared with all possible links, the *average degree* is the mean number of links connected to a node, the *average unit betweenness* is the average value of betweenness centrality, measuring the number of shortest paths passing through a node, normalized to the unit interval, the *diameter* is the greatest shortest path length between any two nodes in the network and *transitivity* measures the local clustering by the empirical probability for a link between two neighbours of a node (for more details see Kolaczyk (2009)).

Holyst (2005) and von Ferber *et al*. (2009); see Table 1). The first PTN, L1, is similar to the German high-speed railway system. In contrast with the network L0 in the application, this PTN consists only of high-speed connections, so it is less detailed and comprises only 319 nodes connected by 416 links. There are few stations of high importance with a large number of links. This network seems to be a good representative for general transportation networks. The other two networks, the bus network of the city of Göttingen (L2) and Athens metro (L3), mainly serve for comparison. The Göttingen bus system is a directed network with 257 stations and 548 connections. It exhibits a strikingly high average degree in comparison with the other regional PTNs. The Athens metro is quite a small network with 51 nodes and 52 links, which results in a relatively low average degree. Additionally, it is noticeable that this network is an extremely centralized network (average unit betweenness 0.18).

3. Network theoretic methods for source detection

In this section, we formalize the source estimation problem and general notation. In what follows, we describe three source estimation approaches that are used in this work.

3.1. Data basis, model assumptions and notation

The goal of our source estimation methods is to find the starting point of general propagation processes on complex networks from a single snapshot about the observed event counts at the network nodes. In this context, we focus on single-source estimation problems.

We denote the underlying network graph as $G = (\mathcal{K}, \mathcal{L})$, which can be specified as a collection of nodes $\mathcal{K} = \{1, \ldots, K\}$ that are connected by direct links from the set $\mathcal{L} = \{(k, l) | k, l \in \mathcal{K}\}$. A path is an ordered sequence of links in the network. When modelling food-borne infectious diseases, the underlying network captures the transportation routes of contaminated food. In the case of train delays, the PTN represents the stations and tracks which are used by trains and on which delays are propagated (see Section 2).

Furthermore, we assume a time-dependent stochastic process $\{X_k(t)\}$ on the network nodes $k \in \mathcal{K}$ characterizing a propagation mechanism in a time range $t = 1, \ldots, T$. The corresponding observations $x_k(t)$ in each node k are collected at different observation times t to find T sequential snapshots of the distribution pattern. In the infectious disease spread context, we assume a susceptible–infected model, where $x_k(t)$ refers to the incidence of infection in a reporting region since onset of the outbreak. During the source estimation analysis in PTNs, $x_k(t)$ corresponds to the relative magnitude of delay observed in a station k within a time slot $[t_0, t]$, i.e. the total magnitude of delay (in minutes) in station k, normalized by the number of trains arriving and departing at this node. For comparison we also analyse data in which $x_k(t)$ corresponds to the number of delayed trains in station k in a time slot $[t_0, t]$. If the spreading onset time t_0 is not known, t_0 can be chosen arbitrarily or as $t_0 = t - 1$. The latter refers to a susceptible–infected–recovered model assumption in the infectious disease context.

3.2. Effective distance median

We generalize the source detection method of Manitz *et al*. (2014), which was originally suggested for the reconstruction of infectious diseases breaking out from an epicentre (see also Brockmann and Helbing (2013)), to general network-based propagation processes, so that it can also be applied to the spread of delays in railway systems.

The key idea assumes that, given a definition of an effective distance, propagation phenomena are spreading in a circular fashion from the origin $k_0 \in \mathcal{K}$ (Manitz *et al*., 2014; Brockmann and

Helbing, 2013). We suggest that the most likely source of a spreading pattern can be simply estimated as the network median as defined in location theory. In contrast with Brockmann and Helbing (2013) and Manitz *et al.* (2014), we obtain more robustness for noisy data by considering not only the process events, but also their observed magnitude. Furthermore, we avoid the estimation of variance, which tends to be unstable, if the number of affected nodes is small.

The effective distance is defined as the effective length of a path γ between any pair of nodes $k, l \in \{1, \ldots, K\}$, which is a combination of the topological length $L(\gamma)$ and the logarithmic path probability $\Pr(\gamma)$, minimized for all possible paths $\gamma \in \Gamma_{kl}$ from origin l to destination k. Thereby, the topological length $L(\gamma)$ is given by the number of links composing the path γ along the nodes $l = k_0, k_1, \ldots, k = k_{L(\gamma)}$. The path probability is the product of the transition probabilities $p_{k_i,k_{i-1}}$ for $i = 1, \ldots, L(\gamma)$ of the corresponding links in the path $\mathcal{L}_\gamma$. A path is considered to be short, if the probability of transiting the path is high, i.e.

$$d_{\text{eff}}(k,l) = \min_{\gamma \in \Gamma_{kl}} [L(\gamma) - \log\{\Pr(\gamma)\}], \qquad \text{for } k, l \in \mathcal{K}. \tag{1}$$

For details on the derivation of this distance see the on-line supplementary material, section 1, as well as Brockmann and Helbing (2013) and Manitz *et al.* (2014). The principal idea underlying source reconstruction is to test different source candidates and to examine the concentricity of the observed pattern on a minimum shortest path tree. This tree is composed of the shortest paths from a candidate k_0 as tree root to all other nodes in the network. Thus, given the effective distance d_{eff}, the source can be reconstructed by minimizing the expected value of the distance $\mu_X(d_{\text{eff}}; k_0, t)$ from the origin k_0 to all other network nodes $k \in \mathcal{K}$ specified by process $X_k(t)$, i.e.

$$\hat{k}_0(t) \in \arg\min_{k_0 \in \mathcal{K}} \mu_X(d_{\text{eff}}; k_0, t). \tag{2}$$

The expected distance $\mu_X(d_{\text{eff}}; k_0, t)$ can then be estimated by the average effective distance $d_{\text{eff}}(k, k_0)$ from source k_0 to all destination nodes k weighted by the observed mean magnitude of delays $x_k(t)$ in node k until time t. Thus,

$$\hat{\mu}_X(d_{\text{eff}}; k_0, t) = \frac{1}{N_x(t)} \sum_{k=1}^{K} x_k(t)\, d_{\text{eff}}(k, k_0), \tag{3}$$

where $N_x(t) = \Sigma_k x_k(t)$ is the total relative delay in the network at time t. Since $\mu_X(d_{\text{eff}}; k_0, t)$ is continuous, we obtain with probability 1 a unique solution.

3.3. Recursive backtracking algorithm for delay propagation

The basic idea of backtracking is the tracing of delays back in time (see, for example, Yamamura *et al.* (2013) and references therein). In comparison with the approach that was described there, the data which are available in our situation are less precise. Hence, we introduce a backtracking method that is adapted to the available data, while explicitly making use of the way that delays spread in a PTN. Given a pattern at time t, let $b_k = 0$ for all $k \in \mathcal{K}$ be a variable counting events as follows.

(a) Consider a node $k \in \mathcal{K}$ which has experienced a relative delay $x_k(t)$.
(b) We look for a node that is adjacent to k with highest relative delay

$$k^* \in \arg\max_{k' \in \mathcal{K}:(k,k') \in \mathcal{L}} x_{k'}(t). \tag{4}$$

(c) If the relative magnitude of delay of this new node is higher than that of the first node ($x_{k^*}(t) \geqslant x_k(t)$), we jump to the new node and repeat ($k := k^*$; go to step (b)).

(d) This process is executed until the current node relative delay magnitude is higher than all of its adjacent nodes ($x_k(t) > \max_{k' \in \mathcal{K}:(k,k') \in \mathcal{L}} x_{k'}(t)$). In this case we increase b_k by 1.

The loop (a)–(d) is repeated for all nodes $k \in \mathcal{K}$ which have experienced delay, i.e. $x_k(t) > 0$. Finally, all nodes $k \in \mathcal{K}$ are ranked by b_k, which gives the number of times that such an iteration ends at each node.

Note that the algorithm assumes that each delayed train carries delays with decreasing magnitude in a particular direction. Aggregating the individual train information results in a spreading in all directions. However, recursive backtracking attempts to invert the mechanism of delay propagation described. As a result, when viewed after the propagation, the path of a delay can be tracked back along increasing delay magnitudes. Hence, backtracking implements this idea of recursively following the path of each delay. We assume that the performance of backtracking improves on a well-defined network, which highlights that the approach is specifically designed for source estimation in PTNs.

3.4. Source estimation approach based on node centrality

In this section, we describe and adapt a simple centrality-based method for comparison. Comin and da Fontoura Costa (2011) suggested that the starting point of a spreading process can be reconstructed as the network node that obtains the highest centrality in the transmission tree. The node centrality is measured by node betweenness normalized by the corresponding node degree. Since in our setting no transmission tree is given, we consider the subgraph that is induced by all nodes which are affected by delays. The subgraph might not be a tree. From this subgraph the node with the highest centrality is estimated to be the source node. Note that it is not possible to compute the betweenness for networks which are composed of fewer than three nodes. In a two-node subgraph, we assign the source estimation to the node that experienced the larger magnitude of event. If the subgraph consists of only one node, this node is estimated as the source node.

4. Delay simulation study

We mimic diverse delay propagation mechanisms on different public transportation networks to obtain different spread patterns. On the basis of these results, we compare the performance of the effective distance median (EDM) source estimation approach and the recursive backtracking algorithm to the adapted centrality-based method of Comin and da Fontoura Costa (2011). We evaluate the performance in dependence of the observation time, analyse the robustness with respect to various PTN structures and assess the influence of background noise.

4.1. Delay simulation setting and performance evaluation

To simulate different delay spreading patterns we use the software toolbox LinTim (Gattermann *et al.*, 2016). We generate an optimized line plan and a timetable for 4 h, consisting of arrival and departure times of all trains at all stations, which represents a timetable as published by railway operators.

In each run of a simulation scenario, we choose one of the stations in the PTN as the source station and introduce 30 sources of delays that are propagated through the system according to a predefined delay management strategy. The delay management strategies differ in their decisions on passenger transfers to be dropped and train sequence reordering and hence lead to different spread patterns. For more details, we refer to the on-line supplementary material, section 1, and Schöbel (2006, 2007) and Schachtebeck and Schöbel (2010).

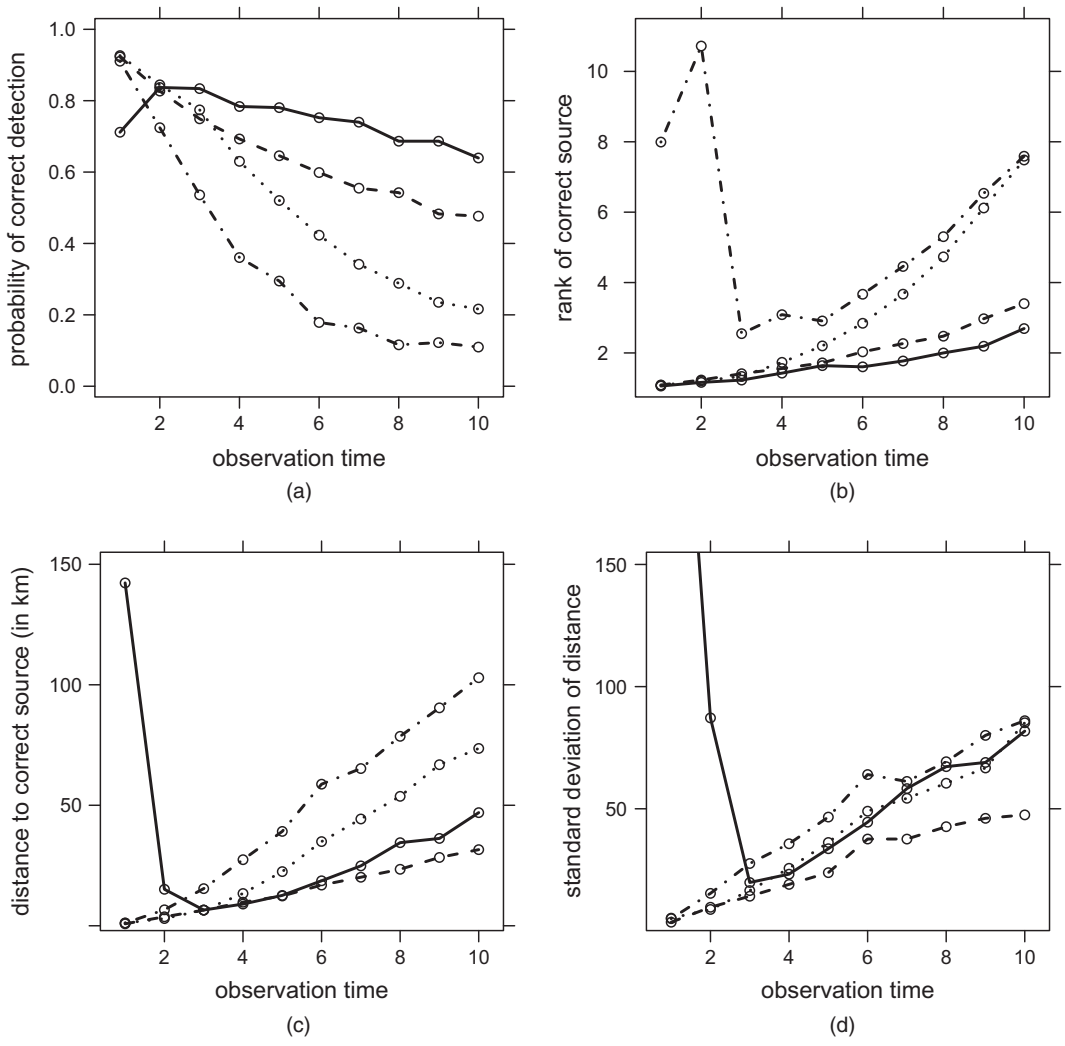

Fig. 1. Detection performance in the standard scenario under consideration of security distances and delay management rule 2 based on L1 (- - - - - -, EDM using relative delay magnitude; ———, EDM using recursive backtracking; · · · · · ·, EDM using counts of delayed trains; - · - · - ·, EDM using the centrality-based method): (a) probability of correct detection; (b) rank of correct source; (c) distance to correct source; (d) standard deviation of distance

The delay dispersal during the 4 h is recorded as 10 sequential snapshots (one every 24 min). The gap between snapshots is chosen to be sufficiently long that the propagation of delays can be observed. For each of the stations in the PTN, we consider the number of delayed trains as well as the total magnitude of delay since the beginning of the dispersal. Typically, we estimate the source considering the relative magnitude of delay, i.e. the total magnitude of delay normalized by the number of trains.

The performance of our source estimation methods is quantified by using four performance measures. The *probability of correct detection* measures the relative number of correct source estimations. From the method-specific ranking of nodes, we report the rank of the correct source as the *rank of correct detection*. Finally, we evaluate the *distance to correct detection*, which is the

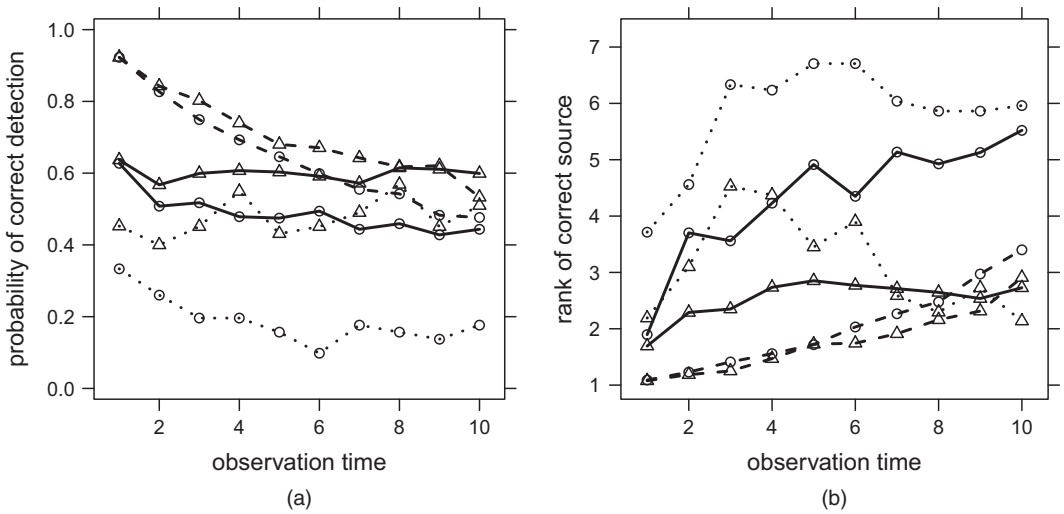

Fig. 2. Comparison of source estimation performance in simulations on various public transportation systems (German railway (– – –, L1), Göttingen bus (———, L2) and Athens metro (· · · · · ·, L3)) (○, EDM; △, backtracking): (a) probability of correct detection; (b) rank of correct source

length of the shortest path between the correct and the detected source node (in kilometres or travel time minutes), and its standard deviation, the *standard deviation of the distance to correct detection*.

4.2. Results

In this section, we compare the methods' performances in various scenarios of delay spreading.

4.2.1. Comparison of source estimation methods

In the first scenario, we simulate 319 delay spread patterns based on a fixed waiting time delay management rule and under consideration of security distances by using the simulated German railway network L1. In this situation, a train waits for transferring passengers from a delayed train only if the delay is below a fixed time, whereas the train sequence is left as planned (rule 2; see the on-line supplementary material, section 2). This is frequently used in practice.

For all source estimation methods, the results reveal decreasing performance as the observation time progresses, whereas the detection rates diverge for the different source estimation methods applied (Fig. 1). This result is intuitive, since the spreading can be regarded as a stochastic process. Hence the delay pattern becomes more complex over time, thereby making source estimation more difficult. We observe that backtracking is the most successful method overall, followed by the EDM method. Note that, for the EDM, better estimation performance is ensured if the more comprehensive relative delay magnitude is taken into account instead of the number of delayed trains. The fact that the backtracking method performs well is to be expected, since the backtracking method is specifically engineered for delay source estimation. The observation that the EDM method with relative delay magnitude performs nearly as well indicates that it can be applied beyond its original context of detecting origins of diseases. Backtracking as well as EDM outperform Comin's centrality-based source estimation method.

The results with respect to the rank of correct detection and distance to correct source are similar. Although backtracking is on average very successful in estimating the correct source,

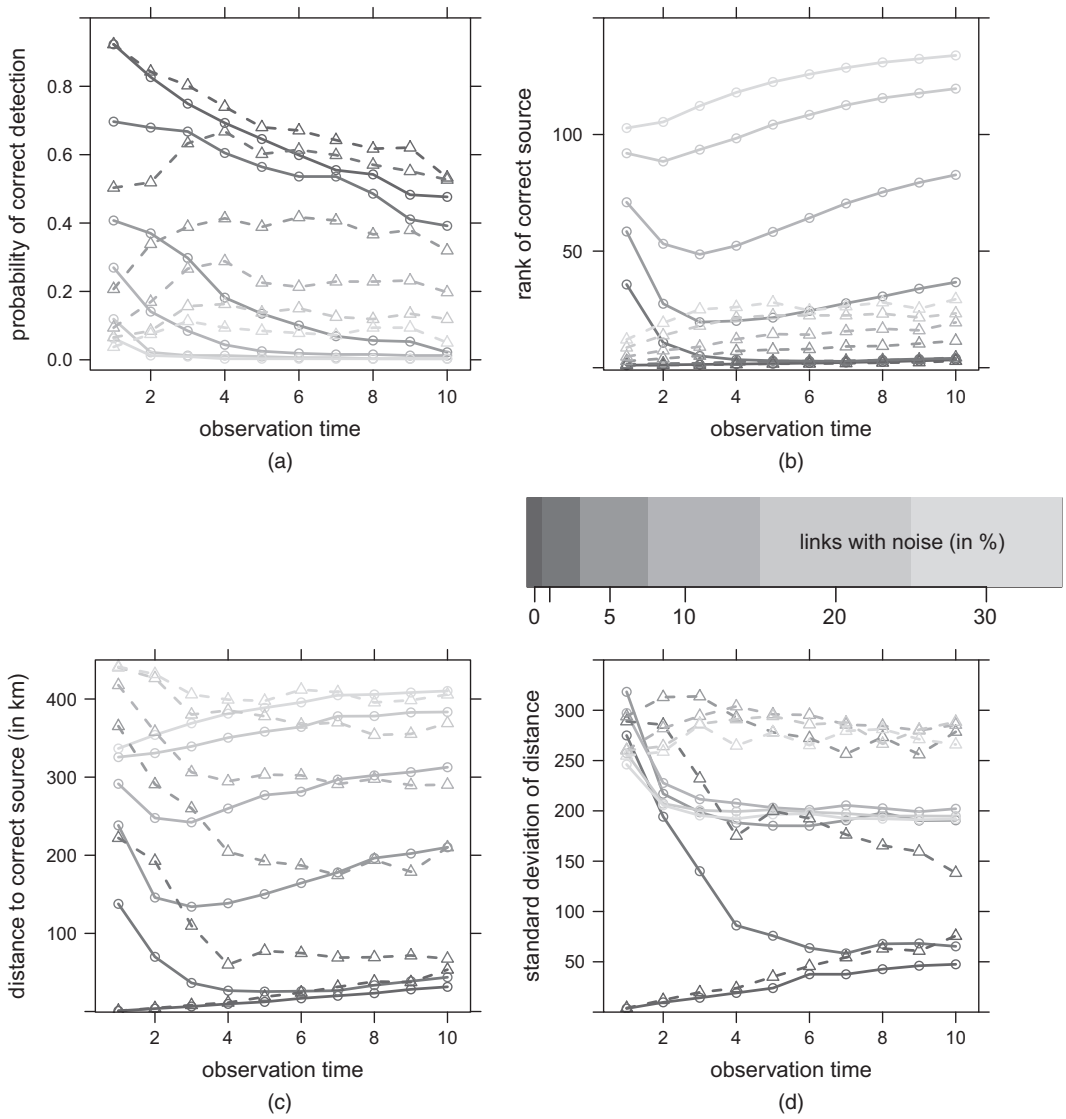

Fig. 3. Comparison of source estimation performance in simulations of scenarios with background noise on L1 (○, EDM; △, backtracking): (a) probability of correct detection; (b) rank of correct source; (c) distance to correct source; (d) standard deviation of distance

it shows large variation in particular at the beginning of the observation period. Altogether, the results show that the methods EDM and backtracking are suitable for source-of-delay estimation.

4.2.2. Network structure

In this analysis, we compare the source estimation performance on the German railway network (L1) with the results from the networks of the Göttingen bus (L2) and the Athens metro (L3). As described in Section 2, the networks have very different layouts. Again, the fixed waiting time delay management rule is used to generate delay spreading patterns.

For both methods, the best source estimation performances can be found on the German railway network, followed by the Göttingen bus network and the Athens metro system (Fig. 2). This evaluation indicates that, for more complex networks, both methods are more reliable. Note that the differences in performance are smaller for backtracking than for the EDM approach. Further, as the structure of the network becomes simpler, the EDM proves to be less effective. This indicates that the EDM is more sensitive to unusual network structures than backtracking. However, it is more important that the methods proposed work well on complex networks, as the spreading patterns are more difficult to observe. We conclude that both methods provide a tool for analysing hidden patterns, particularly for complex networks.

4.2.3. Considering background noise

In realistic spread patterns, we always observe background noise, i.e. small delays triggered at stations which are not considered to be the major source of delay. In this section, we analyse the suggested source estimation approaches with respect to their ability to differentiate between noise and signal delays. Signal delays are those incurred by the source which is to be detected, whereas noise delays include all remaining delays. In addition to the delay pattern in the standard scenario, we add small-to-large delays to the remaining activities. For instance, delaying 1% of the remaining activities corresponds to almost 200 activities (in contrast with 30 source delay activities), where random noise is drawn from an exponential distribution with a mean value of 60 s. Again, the fixed waiting time delay management rule is used to generate the delay spreading patterns on the network L1.

When considering noise, we recognize that estimation performance, in terms of the probability of correct detection, decreases with advancing observation time (Fig. 3). However, in terms of rank of correct detection and distance to correct detection, detection performance improves during the first observation time as the signal comes to the fore. Furthermore, the results clearly indicate that the performance of the EDM and backtracking decreases as the amount of noise increases. As small amounts of noise are imposed, the performances of both methods are still reliable. However, the estimation performance of both methods decreases rapidly for large amounts of noise, because the actual signal is not recognizable anymore. As in the standard scenario, backtracking achieves a slightly better performance in all performance measures except the standard deviation of the distance to correct detection. In this performance measure, the backtracking method shows much higher values, which means that the fluctuation of backtracking is larger than the fluctuation of the EDM.

4.2.4. Further performance results

The influence of time and node centrality on the performance in the standard scenario is further analysed by a generalized additive model for location, shape and scale (Rigby and Stasinopoulos (2005); see the on-line supplementary material, section 5). The results of the model confirm the effect of observation time on source estimation performance and also indicate that node centrality does not have a considerable influence on the estimation performance. Furthermore, even though the probability of detection is found to be slightly lower for the EDM by using the relative delay magnitude in comparison with backtracking, it is not different from a statistical point of view. In terms of distance to correct detection, the performance of EDM estimates by using relative delay magnitude are expected to be closer to the correct source than those from backtracking.

The robustness of the methods with respect to different propagation mechanisms is tested by generating different spreading patterns by using various delay management rules (see the on-line supplementary material, section 3). The results reveal that both the EDM and backtracking

can cope with such different patterns very reliably. We also observe a strong dependence on the total relative delay magnitude in the system, so a larger amount of delays results in lower performance.

Further analyses show that additional knowledge about train or passenger traffic when defining the network improves the source estimation performance only slightly (see the on-line supplementary material, section 4). The source estimation methods perform only slightly worse on unweighted networks than for passenger- and train-weighted networks. Hence, the approaches can be recommended even without knowledge of the network link weighting.

5. Application: train delays on German railway system

In collaboration with the department for Transportation Network Development and Transport Models of the largest German railway operator Deutsche Bahn, three real examples of delay spreading on the German railway system (L0) were selected and analysed by using the source estimation methods developed. We compare the results of the EDM and backtracking with those of the centrality-based source estimation by using the rank of correct detection and distance to correct detection (the travel time in minutes) as defined in Section 4.1. Note that the performance cannot be measured in terms of the probability of correct detection since for every observation time there is only one estimation.

The railway network consists of 1049 stations that are connected by 3484 links, including regional connections (for more details see Table 1). However, for this analysis only the delays for high-speed trains are available. As a result, we have information on only delays for a subnetwork with fewer than 300 nodes (depending on the example). Because of confidentiality, we do not provide the names of the stations, but their degrees and betweenness centralities c_D and c_B, respectively. After 6 a.m., the data are aggregated to state the relative magnitudes of delay with respect to 30-min time slots day long.

5.1. Source detection results

In the first example, the delays originate in station x.152 ($c_D = 144; c_B = 131\,897$), where damage was recorded at 6.25 a.m. We have delay data for 289 stations. Using the delay patterns from 3 h after the damage report, the EDM method performs well and locates the actual source (Fig. 4 and the on-line supplementary material, section 6). On the basis of data recorded the first 3 h after the damage report, the EDM estimates station x.105 ($c_D = 113; c_B = 115020$) to be the source. This station is 126 min from the actual source station. This station is known to introduce large disturbances in the system. In comparison, the recursive backtracking method identifies the true origin or a station nearby ($c_D \in [15; 144]; c_B \in [1068; 131\,897]$) on the basis of all delay patterns recorded within 4 h of the damage report. However, the estimation is less stable in its reliability compared with other methods. The estimations based on delay patterns recorded later than 4 h after the damage report detect sources in an area which is known to be a large tunnel construction site. The centrality-based approach never identifies the origin of delays but always ranks it within the top 10 throughout the day. However, the method finds stations which are smaller and more than 200 km from the correct detection ($c_D \in [5; 18]; c_B \in [4245; 15\,237]$).

In another example, delays are reported in 297 stations. The signal delays are caused at station x.95004823 ($c_D = 36; c_B = 14671$) from 10.30 a.m. because of an electrical failure in a switch tower. The relative performances of the methods are similar to those in the first example. On the basis of the first observation times the EDM identifies a large high influence station and about 3 h after the damage the actual source is identified correctly. Again, the backtracking

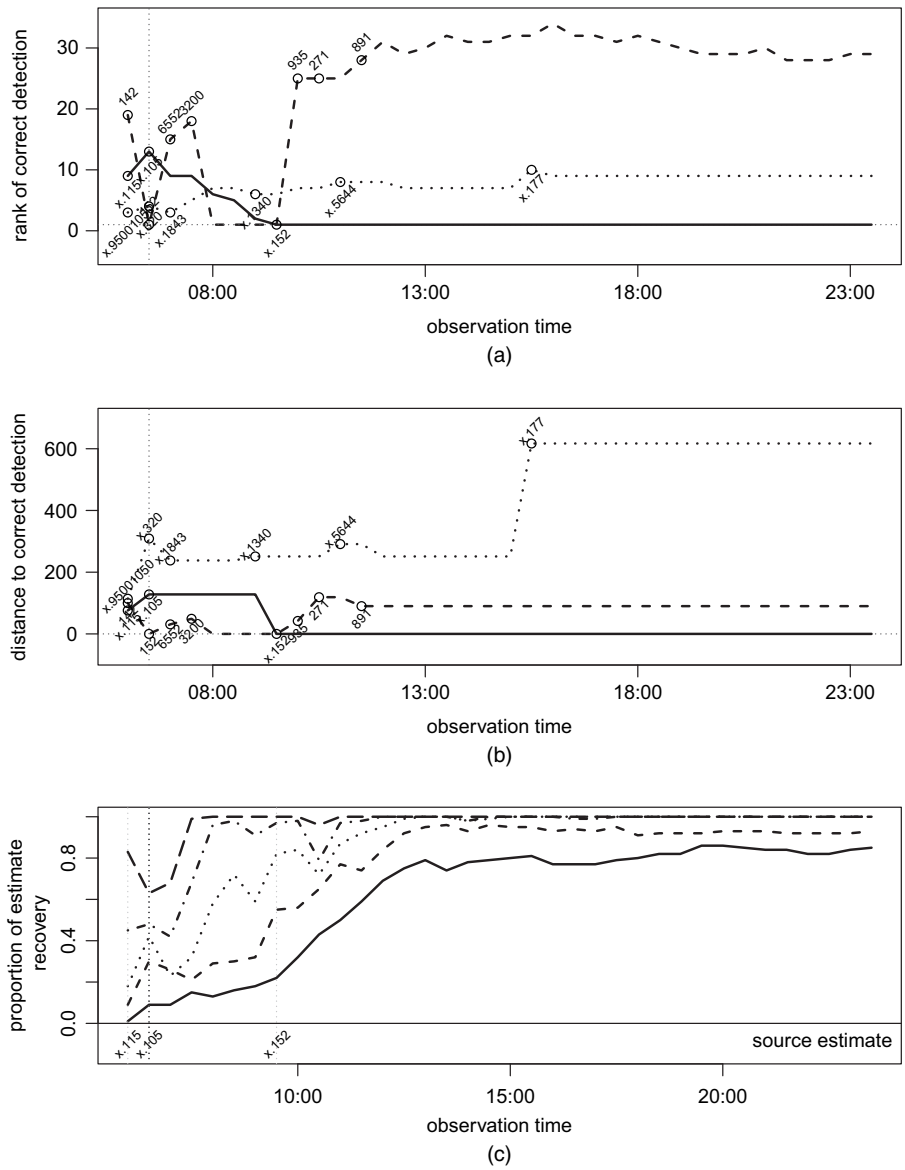

Fig. 4. Source estimation in the Deutsche Bahn delay example 1 with delay source in station x.152 at 6.25 a.m. on L0: (a) rank of correct detection (———, EDM; – – –, backtracking; · · · · · ·, centrality-based method); (b) distance to correct detection; (c) uncertainty assessment (- - - - - -, $p = 0.1$; · · · · · ·, $p = 0.3$; - · - · - ·, $p = 0.5$; – – –, $p = 0.7$; ———, $p = 0.9$)

locates the region of the correct source shortly after the damage report but the estimations based on patterns 4 h after the damage report lead to a station about 81 km from the real source. The centrality-based approach detects only sources which are much further compared with the previous methods. Finally, we analysed a delay spread pattern (recorded data for 299 stations) originating from station x.105 ($c_D = 113; c_B = 115020$) at 7.45 a.m. Also, in this example, the relative performances of the different approaches hold (see the on-line supplementary material, section 6).

Altogether, the application of the methods suggested to the three real world examples of delay spreading gives reasonable results and shows the general applicability of the methods on real data. The EDM method seems to be a robust approach as it reliably identifies the origin on the basis of delay patterns recorded a few hours after the damage report. In contrast, backtracking is a highly sensitive approach. Backtracking determines the area of the source on the basis of patterns recorded shortly after the damage report but has difficulties settling on a particular station. As a result, the method is quite unstable. The backtracking method's performance decreases when applied to delay spreading patterns which are recorded a few hours after the damage report. The centrality-based source estimation is the least reliable approach.

Considering the potential effect of source node centrality, we make the following observations: In all examples, we recognize that station x.105 is likely to be detected, in particular by the EDM. This station is a node of high centrality and is known to cause many disturbances in the German railway system. However, in the second example, the EDM identifies the actual origin ($c_D = 36; c_B = 14671$) even though a neighbouring node has much higher centrality (x.93244; 9 min travel time; $c_D = 55; c_B = 30189$).

5.2. Uncertainty assessment via subsampling for effective distance median estimates

The source estimation methods presented result in estimates without quantifying their uncertainty. In what follows, we show the construction of an assessment of uncertainty for EDMs by using a subsampling procedure. This is inspired by the idea of variation estimation with delete d jackknife resampling, which is a linear approximation of the bootstrap for estimators that are not 'smooth' (Efron and Tibshirani, 1994).

We create subsamples of individual trains and their punctuality by using a sampling proportion $p \in [0, 1]$. After aggregating the data, we apply the source estimation approach. Using this result, we deduce the relative frequency of how often the source estimate that is obtained with the complete data set can be recovered by source estimation based on the subsample. Thus, the uncertainty of the estimate is assessed by the *proportion of estimate recovery*. Since we use a subsampling technique, we underestimate the true uncertainty of estimation. Thus, we deduce only an upper bound for the probability of estimate recovery. However, for a fixed subsampling proportion p, the results are comparable with advancing observation time. Furthermore, the proportion of estimate recovery can be used to construct confidence sets of nodes that are likely to be the source of spreading.

As an example, the procedure is applied to the case of the EDM and on the first example of delay spreading on the German railway system by using different sample proportions; see Fig. 4(c). We use 100 resamplings of individual trains and their punctuality information. For incorrect source estimate x.105 between 6.30 a.m. and 9 a.m., the proportion of estimate recovery fluctuates for all sampling proportions considerably. After identifying the correct origin in node x.152 at 9.30 a.m., the proportions of estimate recovery increases more steadily. Note that the proportions of estimate recovery for $p = 0.7$ and $p = 0.9$ do not differ considerably, so lower sampling proportions seem to be more informative. Using the assessment of uncertainty on a specific case of application, where the true origin is not known, we suggest selecting a fixed proportion for sampling depending on the signal strength presumed.

6. Conclusion and discussion

We propose two source estimation approaches that are conceptually simple and computationally efficient: EDM and recursive backtracking. The EDM estimates the source on the basis of the

effective distance, whereas recursive backtracking is specifically designed for the estimation of sources of delays in public transport.

Both methods show good performance with differing strengths and weaknesses when applied to a simulation study as well as to a real example from the largest German railway operator Deutsche Bahn. In the simulation study, backtracking has somewhat better performance but has the drawback of larger variations compared with the EDM approach. In the real examples, the EDM source estimation is a robust method for source detection, which steadily identifies the actual origin on the basis of the delay spread patterns that are recorded a few hours after a damage report. Recursive backtracking complements the EDM by being a sensitive method, which locates the area of origin shortly after the damage report and for the first observation times, but the estimation loses accuracy as the observation time advances. Note that the observations in the simulation study are not much different from the findings based on the real examples. In the simulation study with background noise, we observe that the performance of all methods increases for the first observation times and decreases after that. In the real examples, the performance of the EDM steadily improves as the observation time advances. We speculate that this improvement is due to a larger quantity of data and an increased ability to distinguish noise and signal. Hence, the results from the real examples can be seen as a further extension of the results from the simulation study. The analysis of the simulation study and the real examples both show that the EDM and backtracking are effective for source detection, whereas the centrality-based method proves to be inferior compared with both.

Although the EDM approach was originally developed for application to infectious disease propagation, the method proves to be applicable to delay spreading in train networks. We also analysed whether backtracking is applicable to other spreading processes. For comparison, we tested the recursive backtracking approach on food-borne disease spreading data from the enterohaemorrhagic *Escherichia coli*–haemolytic uraemic syndrome outbreak. This application assumed a food dispersal network approximated by the gravity model of trade (see Manitz *et al.* (2014)). The backtracking results show that its source estimate is very inaccurate; there were no examples of cases in which the source was determined correctly. Backtracking exhibits a preference for estimating source nodes which are better connected and have higher incidence of infection compared with the correct source. We conclude that this method is designed for situations in which the source node constantly introduces new delays in the network. Accordingly, its performance decreases when the source node stops introducing new delays in the system. The method is not limited to circular spreading patterns, but in most cases it determines the source as the node with the highest relative delay.

Note that our approaches are designed for the estimation of a single source. This assumption can seem very restrictive. However, Zang *et al.* (2014) recently suggested a method that converts a multiple-source location problem into a number of single-source detection problems. This approach is based on the decomposition of the network graph into a number of subgraphs by applying simple community cluster algorithms. Initial simulations show promising results (see the on-line supplementary material, section 7), and it would be interesting to investigate the performance in an extensive simulation study. Furthermore, we discussed only sources of delays which originate from stations, but the methods can also be adjusted to deal with sources of delays originating from failures between stations.

The results encourage the application of these methods (in particular of the EDM approach) to various source estimation problems, e.g. the roots of delays in supply chains processes (Giannakis and Louis, 2011), initial failure detection during blackouts in power grids (Crucitti *et al.*, 2004), the origin of computer virus attacks in the Internet (Shah and Zaman, 2010), the source of invasive species in ecology (Stevenson *et al.*, 2012), the beginning of rumour or misinformation

in social networks (Shah and Zaman, 2012) and also the reconstruction of the epicentre of outbreaks of infectious disease (Pinto *et al.*, 2012). Knowledge of the origin of a propagation process empowers one to prevent further spreading.

7. Data accessibility

The network data of the simulation studies (L1, L2 and L3) were obtained from LinTim (Gattermann *et al.*, 2016) and are freely accessible. LinTim was used also to generate the delay patterns for our simulation. The performance evaluation and statistical network analysis were conducted with the statistical software package R (R Core Team, 2014). The related R package `NetOrigin` containing the implementation for the methods presented in this paper is available on the Comprehensive R Archive Network (Manitz and Harbering, 2016). Because of a confidentiality agreement, we cannot make the Deutsche Bahn delay example data (L0) available.

Acknowledgements

We thank the *Deutsche Bahn* Mobility Logistics AG, *Verkehrsnetzentwicklung und Verkehrsmodelle*, for the contribution of expert knowledge and making real data examples available. Special thanks go to Ingmar Schüle for the productive collaboration. This work was supported by the German Research Foundation, Research Training Group 1644 'Scaling problems in statistics', the Simulation Science Center Clausthal–Göttingen and the Alexander von Humboldt Foundation. Furthermore, we thank the reviewers for their helpful comments and Christopher Gruber for checking the manuscript.

References

Adar, E. and Adamic, L. (2005) Tracking information epidemics in blogspace. In *Proc. Int. Conf. Web Intelligence*, pp. 207–214. New York: Institute of Electrical and Electronics Engineers.

Brockmann, D. and Helbing, D. (2013) The hidden geometry of complex, network-driven contagion phenomena. *Science*, **342**, 1337–1342.

Comin, C. H. and da Fontoura Costa, L. (2011) Identifying the starting point of a spreading process in complex networks. *Phys. Rev.* E, **84**, article 056105.

Crucitti, P., Latora, V. and Marchiori, M. (2004) A topological analysis of the Italian electric power grid. *Physica* A, **338**, 92–97.

Efron, B. and Tibshirani, R. J. (1994) *An Introduction to the Bootstrap*. New York: Chapman and Hall.

von Ferber, C., Holovatch, T., Holovatch, Y. and Palchykov, V. (2009) Public transport networks: empirical analysis and modeling. *Eur. Phys. J.* B, **68**, 261–275.

Fioriti, V., Chinnici, M. and Palomo, J. (2014) Predicting the sources of an outbreak with a spectral technique. *Appl. Math. Sci.*, **8**, 6775–6782.

Gattermann, P., Harbering, J., Schiewe, A. and Schöbel, A. (2016) LinTim—integrated optimization in public transportation. (Available from `http://lintim.math.uni-goettingen.de/`.)

Giannakis, M. and Louis, M. (2011) A multi-agent based framework for supply chain risk management. *J. Purchsng Supply Mangmnt*, **17**, 23–31.

Goerigk, M., Schachtebeck, M. and Schöbel, A. (2013) Evaluating line concepts using travel times and robustness: simulations with the Lintim toolbox. *Publ. Transprt*, **5**, 267–284.

Jiang, J., Wen, S., Yu, S., Xiang, Y., Zhou, W. and Hossain, E. (2014) Identifying propagation sources in networks: state-of-the-art and comparative studies. *IEEE Communs Surv. Tutorls*, **17**, in the press.

Kolaczyk, E. D. (2009) *Statistical Analysis of Network Data: Methods and Models*. New York: Springer.

Lappas, T., Terzi, E., Gunopulos, D. and Mannila, H. (2010) Finding effectors in social networks. In *Proc. 16th Special Interest Group on Knowledge Discovery and Data Mining Int. Conf. Knowledge Discovery and Data Mining*, pp. 1059–1068. New York: Association for Computing Machinery Press.

Manitz, J. and Harbering, J. (2016) NetOrigin: origin information for propagation processes in complex networks. *R Package Version 0-2*. Georg-August-Universität Göttingen, Göttingen. (Available from `https://CRAN.R-project.org/package=NetOrigin`.)

Manitz, J., Kneib, T., Schlather, M., Helbing, D. and Brockmann, D. (2014) Origin detection during food-borne disease outbreaks—a case study of the 2011 EHEC/HUS outbreak in Germany. *PLOS Curr. Outbrks*, **1**, 1–31.
Pinto, P. C., Thiran, P. and Vetterli, M. (2012) Locating the source of diffusion in large-scale networks. *Phys. Rev. Lett.*, **109**, article 068702.
Plöchinger, S. and Jaschensky, W. (2013) So verspätet ist die Bahn in Ihrer Stadt, auf Ihrer Strecke. In *Sueddeutsche Zeitung*. München. (Available from `http://www.sueddeutsche.de/reise/verspaetungs-atlas-so-verspaetet-ist-die-bahn-in-ihrer-stadt-auf-ihrer-strecke-1.1651455`.)
Prakash, B. A., Vreeken, J. and Faloutsos, C. (2012) Spotting culprits in epidemics: how many and which ones? In *Proc. 12th Int. Conf. Data Mining*, pp. 11–20. New York: Institute of Electrical and Electronics Engineers.
R Core Team (2014) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507–554.
Schachtebeck, M. and Schöbel, A. (2010) To wait or not to wait—and who goes first?: Delay management with priority decisions. *Transprtn Sci.*, **44**, 307–321.
Schöbel, A. (2006) *Optimization in Public Transportation: Stop Location, Delay Management and Tariff Planning from a Customer-oriented Point of View*. New York: Springer.
Schöbel, A. (2007) Integer programming approaches for solving the delay management problem. In *Algorithmic Methods for Railway Optimization*, pp. 145–170. Berlin: Springer.
Shah, D. and Zaman, T. (2010) Detecting sources of computer viruses in networks: theory and experiment. In *Proc. SIGMETRICS Int. Conf. Measurement and Modeling of Computer Systems*, pp. 203–214. New York: Association for Computing Machinery.
Shah, D. and Zaman, T. (2012) Rumor centrality: a universal source detector. In *Proc. SIGMETRICS/PERFORMANCE Jt Int. Conf. Measurement and Modeling of Computer Systems*, pp. 199–210. New York: Association for Computing Machinery.
Sienkiewicz, J. and Holyst, J. (2005) Statistical analysis of 22 public transport networks in Poland. *Phys. Rev.* E, **72**, article 046127.
Stevenson, M. D., Rossmo, D. K., Knell, R. J. and Le Comber, S. C. (2012) Geographic profiling as a novel spatial tool for targeting the control of invasive species. *Ecography*, **35**, 704–715.
Yabuki, H., Ageishi, T. and Tomii, N. (2015) Mining the cause of delays in urban railways based on association rules. In *Proc. 13th Conf. Advanced Systems in Public Transport*. (Available from `www.caspt.org/proceedings/paper68.pdf`.)
Yamamura, A., Koresawa, M., Adachi, S. and Tomii, N. (2013) Identification of causes of delays in urban railways. In *Computers in Railways*, vol. XIII, *Computer System Design and Operation in the Railway and Other Transit Systems* (eds C. A. Brebbia, N. Tomii, J. M. Hera, B. Ning and T. Tzteropoulos), pp. 403–414. Southampton: WIT.
Zang, W., Zhang, P., Zhou, C. and Guo, L. (2014) Discovering multiple diffusion source nodes in social networks. *Proc. Comput. Sci.*, **29**, 443–452.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary materials: Source estimation for propagation processes on complex networks with an application to delays in public transportation systems'.