Eur. Phys. J. B (2016) 89: 61
DOI: 10.1140/epjb/e2016-60663-0

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities*

Ingo Scholtes[a], Nicolas Wider, and Antonios Garas

ETH Zürich, Chair of Systems Design, Weinbergstrasse 56/58, 8092 Zürich, Switzerland

Received 10 August 2015 / Received in final form 18 January 2016
Published online 2 March 2016 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2016

Abstract. Despite recent advances in the study of temporal networks, the analysis of time-stamped network data is still a fundamental challenge. In particular, recent studies have shown that correlations in the *ordering of links* crucially alter *causal topologies* of temporal networks, thus invalidating analyses based on static, time-aggregated representations of time-stamped data. These findings not only highlight an important dimension of complexity in temporal networks, but also call for new network-analytic methods suitable to analyze complex systems with time-varying topologies. Addressing this open challenge, here we introduce a novel framework for the study of *path-based centralities* in temporal networks. Studying betweenness, closeness and reach centrality, we first show than an application of these measures to time-aggregated, static representations of temporal networks yields misleading results about the actual importance of nodes. To overcome this problem, we define path-based centralities in *higher-order aggregate networks*, a recently proposed generalization of the commonly used static representation of time-stamped data. Using data on six empirical temporal networks, we show that the resulting higher-order measures better capture the true, *temporal* centralities of nodes. Our results demonstrate that higher-order aggregate networks constitute a powerful abstraction, with broad perspectives for the design of new, computationally efficient data mining techniques for time-stamped relational data.

1 Introduction

The network perspective has provided valuable insights into the structure and dynamics of numerous complex systems in nature, society and technology. However, most of the complex systems studied from this perspective are not static, but rather exhibit time-varying interaction topologies in which elements are only linked to each other at specific times or during particular time intervals. While the *topological* characteristics resulting from which elements are linked to which other elements have been studied extensively, the importance of the additional *temporal* dimension resulting from *when* these links occur has become clear only recently. And despite an increasing volume of research, its full impact on the properties of complex systems and on the evolution of dynamical processes still eludes our understanding [1,2].

Addressing this open issue, different strands of research have focused on the question how different types of temporal characteristics of complex networked systems – such as the activation times of nodes, the inter-event times between links, the duration and/or concurrency of interactions, or the order in which these interactions occur – affect the properties of temporal networks as well as dynamical processes evolving on them. For a couple of systems, it was shown that inter-event times follow heavy-tailed distributions which in turn significantly influence the speed of processes like spreading and diffusion [3–13].

Apart from the *timing* of interactions, the *order* in which these interactions occur is another important characteristic of temporal networks. Not only does the ordering of interactions crucially affect causality in temporal networks, it has also been shown to dramatically shift the evolution of dynamical processes compared to what we would expect based on a static, time-aggregated perspective [14–18]. Some of these works have further taken a modeling perspective, highlighting that real-world temporal network data exhibit non-Markovian characteristics in the sequence of links which are not in line with the Markovianity assumption that is (implicitly) made when studying static representations of time-varying complex networks. Neglecting these non-Markovian characteristics not only leads to wrong results about dynamical processes, it also leads to wrong centrality-based rankings of nodes, as well as misleading results about community structures [16–18].

The main reason why an analysis of static, time-aggregated networks yields misleading results about the properties of temporal networks is that the ordering of links can alter path structures in temporal networks compared to what we would expect based on their static topology. Precisely, in static networks the presence of two

* Contribution to the Topical Issue "Temporal Network Theory and Applications", edited by Petter Holme.

[a] e-mail: ischoltes@ethz.ch

links (a,b) and (b,c) connecting nodes a to b and b to c necessarily implies that a path from a via b to c exists. However in a temporal network, for a to be able to influence c the link (a,b) must occur *before* the link (b,c) and thus the presence of a path depends on the ordering of links. This simple example highlights that the mere ordering of links in temporal networks can introduce an additional temporal-topological dimension that can neither be understood from the analysis of static, time-aggregated representations, nor from the analysis of inter-event times or node activity distributions [15].

Highlighting the important consequences introduced by the specific ordering of links in real-world temporal networks, in this article we study how this ordering affects path-based centrality measures in temporal networks. The main contributions of our work are as follows:

1. Building on the concept of time-respecting paths with a maximum time difference between consecutive links as previously discussed in references [1,19], we introduce three different notions of path-based temporal node centralities which emphasize the additional *temporal-topological* dimension that is introduced due to the ordering of links in temporal networks. In particular, we formally define temporal variations of *betweenness*, *closeness* and *reach* centrality and demonstrate how they can be computed based on the topology of shortest time-respecting paths emerging in temporal networks.
2. Calculating these temporal centrality measures for six empirical data sets, we quantify to what extent a ranking of nodes based on temporal centralities coincides with a ranking of nodes based on the same measures, however calculated based on the corresponding static, time-aggregated networks. From our results we conclude that, possibly due to non-Markovian characteristics previously highlighted in references [15,17], a static analysis of node centralities yields misleading results about the importance of nodes with respect to time-respecting paths.
3. Generalizing the usual time-aggregated static perspective on temporal networks, we further develop the second-order time-aggregated representations introduced in reference [17], obtaining higher-order time-aggregated representations which can be conveniently analyzed using standard network-analytic methods. Notably, despite being static representations of temporal networks, we show that these higher-order representations allow to incorporate those order correlations that have been shown to influence the causal topologies of temporal networks.
4. We finally define generalizations of static betweenness, closeness and reach centrality based on a second-order aggregate representation of temporal networks. Using six data sets on temporal networks, we show that these second-order generalizations of centralities constitute highly accurate approximations for the true temporal centrality of nodes calculated based on the detailed time-respecting path structures in temporal networks.

The remainder of this article is structured as follows: in Section 2 we first introduce basic concepts such as our notion of temporal networks, time-aggregated and time-unfolded representations of temporal networks, as well as time-respecting paths with maximum time differences between consecutive links. In Section 3 we introduce the framework of higher-order time-aggregated networks, a simple abstraction of temporal networks that takes into account the statistics of time-respecting paths up to a given length. In Section 4 we finally define three temporal centrality measure which account for the temporal-topological characteristics introduced by the shortest time-respecting path structures in real-world temporal networks. Comparing the importance of nodes according to (i) temporal centralities, (ii) centralities calculated based on a commonly used static, time-aggregated representation, and (iii) second-order centralities calculated based on a static, second-order time-aggregated representation, we show that higher-order aggregate networks provide interesting perspectives for the analysis of temporal networks. We finally conclude our article by a summary of key contributions and a discussion of open issues and future work.

2 Temporal networks and time-respecting paths

In this section, we formally introduce the basic concepts and definitions used throughout our work. In particular, we define the notion of a *temporal network* used throughout this article, as well as *time-respecting paths* which are the basis for the notions of *distances* and *path-based centralities* in temporal networks which will be used in subsequent sections.

2.1 Temporal, time-aggregated and time-unfolded networks

We define a temporal network $G^T = (V, E^T)$ as a tuple consisting of a set of nodes V and a set $E^T \subseteq V \times V \times [0,T]$ of time-stamped links $(v,w;t) \in E^T$ for an observation period $[0,T]$. Importantly, we assume *discrete* time stamps $t \in [0,T]$ and time-stamped links $(v,w;t)$ which indicate the presence of the link (v,w) at time t. This "instantaneous" definition particularly does not allow links to be assigned a *duration*, i.e. we cannot directly assign links a time interval during which they exist. However, we can nevertheless represent links that persist for some time interval $[t_{start}, t_{end}]$ by assuming some small unit of discrete time Δt and adding multiple time-stamped links $(v,w;t)$ at time stamps $t = t_{start}, t_{start}+\Delta t, t_{start}+2\Delta t, \ldots, t_{end}$. These assumptions naturally lend themselves to real-world time-stamped data sets, which are typically obtained based on some sort of *sampling*, whose sampling frequency defines the smallest unit of time Δt.

For illustrative purposes it is often useful to be able to visualize temporal networks. Throughout this article,

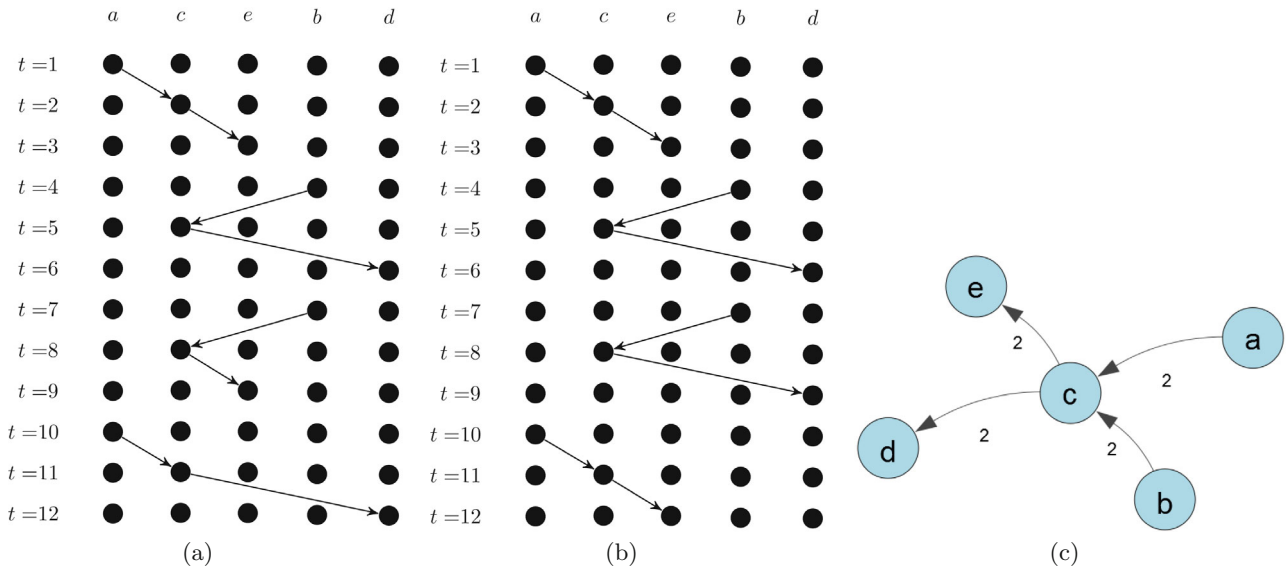

Fig. 1. Time-unfolded and weighted static, time-aggregated representation of two temporal networks G_1 (a) and G_2 (b). (c) Weighted, time-aggregated representation of both G_1 and G_2.

we will use so-called *time-unfolded networks*, a simple and intuitive static representation of temporal networks which, in different variants, has been used in a number of previous works [15,17,20,21]. The key idea of this two-dimensional static representation is to arrange all nodes on a horizontal dimension, while unfolding time to an additional vertical dimension as illustrated in Figure 1. For an observation period $[0, \ldots, T]$ and a given Δt we can then add *temporal copies* of all nodes for all possible time steps $k\Delta t$ (for $k = 0, 1, \ldots$). For simplicity, in the following we assume $\Delta t = 1$, which allows us to denote the temporal copies of a node v as $v_t, v_{t+1}, v_{t+2}, \ldots$. The main benefit of this construction is that it allows us to represent a time-stamped link $(v, w; t)$ by means of a static link (v_t, w_{t+1}) connecting the temporal copies v_t and w_{t+1} of node v and node w, respectively. The intuition behind this notation is that a quantity residing at node v at time t can move to node w via a time-stamped link $(v, w; t)$, arriving there at the next time step $t + 1$. Two simple examples for time-unfolded static representations of two different temporal networks with five nodes and eight time-stamped links are shown in Figures 1a and 1b.

Despite the recent development of methods to study temporal networks, the most wide-spread way to study time-stamped network data is to aggregate all time-stamped links into a static, *time-aggregated network* $G = (V, E)$. This means that, given a temporal network $G^T = (V, E^T)$, two nodes $v, w \in V$ are connected in the static network whenever a time-stamped link exists at *any* time stamp, i.e., $(v, w) \in E$ if $(v, w; t) \in E^T$ for any $t \in [0, T]$. Additional information about the statistics of time-stamped links in the underlying temporal network can be preserved by considering a *weighted time-aggregated* network, in which weights $\omega(v, w)$ indicate the number of times time-stamped links $(v, w; t)$ have been active during the observation period. I.e., we consider a weighted time-aggregated network with a weight function $\omega : E \to \mathbb{N}$ defined as:

$$\omega(v, w) := \left|\left\{t \in [0, T] \,|(v, w; t) \in E^T\right\}\right|.$$

Figure 1c shows the weighted, time-aggregated networks corresponding to the two temporal networks shown in Figures 1a and 1b. These simple examples highlight the important fact that *different* temporal networks are consistent with the *same* weighted, time-aggregated network. This is due to the fact that in the time-aggregated network we lose all information on both the timing and the ordering of links in the temporal network.

2.2 Time-respecting paths

Importantly, both the timing and the ordering of links influence path structures in temporal networks. In particular, in the context of temporal networks we must consider *time-respecting paths*, an extension of the concept of paths in static network topologies which additionally respects the timing and ordering of time-stamped links [1,19,22]. For the remainder of this paper, we define a time-respecting path between a source node v and a target node w to be any sequence of time-stamped links

$$(v_0, v_1; t_1), (v_1, v_2; t_2) \ldots, (v_{l-1}, v_l; t_l)$$

such that $v_0 = v, v_l = w$ and the sequence of time-stamps is increasing, i.e. $t_1 < t_2 \ldots < t_l$. The latter condition on the ordering of links is particularly important since it is a necessary condition for causality. This means that for any temporal network a node a is able to influence node c based on two time-stamped links (a, b) and (b, c) only if link (a, b) has occurred *before* link (b, c). A simple example for a time-respecting path $(a, c; 1)$, $(c, d; 5)$ can be seen in

Figure 1a, where the time-unfolded representation of the temporal network G_1 is illustrated.

At this point, it is important to note that, different from the usual notion of paths in static networks, the question whether a time-respecting path exists between two nodes requires to specify a *start time* $t_0 \leq t_1$. In the example of Figure 1a we observe a time-respecting path $(a, c; t_1 = 1), (c, d; t_2 = 5)$ between node a and d, which can only be taken if we consider paths starting at node a at time $t_0 = 1$. If instead we were to ask for a time-respecting path between a and d starting at node a at time $t_0 = 5$, our only choice would be the path $(a, c; 10), (c, d; 11)$.

2.3 Time-respecting paths with a maximum time difference

In the definition of a time-respecting path above, we have required that the sequence of time stamps of the links constituting the path must be increasing. Clearly, this condition is rather weak since it makes no assumptions whatsoever about the time difference between two consecutive time-stamped links on a time-respecting path. As such, for the mere existence of a time-respecting path in a temporal network evolving over a period of years, it is actually not important whether the time difference between two consecutive links is a few seconds or a few years.

However, we typically study time-respecting path structures because they constitute the substrate for the evolution of dynamical processes which have intrinsic time scales that are much smaller than the period during which we observe a temporal network. In the study of time-respecting paths, it is thus often reasonable to impose a *maximum time difference* δ, i.e. we limit the temporal gaps between two consecutive time-stamped links that are considered to contribute to a time-respecting path to a maximum of δ [1,19]. In this case, rather than requiring a mere increasing sequence of time-stamps, we demand that the condition $0 < t_{i+1} - t_i \leq \delta$ must be fulfilled for all $i = 1, \ldots, l-1$. For a maximum time difference of $\delta = 1$, we thus limit ourselves to the study of time-respecting paths for which all time-stamped links occur at immediately consecutive time stamps. As another limiting case, we can consider $\delta = \infty$, which means that we impose no further condition apart from the requirement that the sequence of time stamps of links on a time-respecting path is increasing. Revisiting the example of Figure 1a, we observe that the time-respecting path $(a, c; 1), (c, d; 5)$ only exists if we allow for a maximum time difference $\delta = 4$, while for all $\delta < 4$ the only time-respecting path between the nodes a and d is $(a, c; 10), (c, d; 11)$.

2.4 Shortest and fastest time-respecting paths

Let us now formally define the length of time-respecting paths in a temporal network, which will allow us to define the notion of *shortest time-respecting* paths used throughout our work. Due to the additional temporal dimension, the length of a time-respecting path

$$(v_0, v_1; t_1), \ldots, (v_{l-1}, v_l; t_l)$$

can be studied both from a topological and a temporal perspective. Following the usual terminology, we call the number l of time-stamped links on a time-respecting path the (topological) *length* of the path. We further call the time difference $t_l - t_1 + 1$ the *duration* of the path. Here the increment by one accounts for the duration of the final link $(v_{l-1}, v_l; t_l)$, i.e. for the fact that any process starting at node v_0 at time t_1 will only reach node v_l at time t_{l+1}.

Having defined both the length and duration of time-respecting paths, it is now trivial to define the *shortest time-respecting path* between two nodes v and w as the time-respecting path with the smallest (topological) length. In analogy, we define the *fastest time-respecting path* as the time-respecting path with the smallest (temporal) duration. Following our previous comment about the necessity to define a start time t_0 for a time-respecting path, it is clear that the shortest or fastest time-respecting path can only be found unambiguously with respect to a given start time t_0, i.e. at different times during the evolution of a temporal network the same pair of nodes can be connected by different shortest or fastest time-respecting paths.

2.5 Transitivity of paths in static and temporal networks

Let us conclude this preliminary section by highlighting important differences between paths in static networks compared to time-respecting paths in temporal networks, that result from the ordering and timing of links. Let us first highlight that paths in static networks are *transitive*. This means that from the presence of two paths $(v_0, v_1), \ldots, (v_{k-1}, v_k)$ and $(v_k, v_{k+1}), \ldots, (v_{l-1}, v_l)$ between v_0 and v_k and between v_k and v_l respectively, we can conclude that a path $(v_0, v_1), \ldots, (v_{l-1}, v_{v_l})$ between nodes v_0 and v_l necessarily exists[1]. This transitivity has the important mathematical consequence that the entries in the kth power A^k of the adjacency matrix A of a static network topology count all possible paths of length k between all possible pairs of nodes. Furthermore, transitivity of paths is the basis for a wealth of *algebraic network-analytic methods* such as spectral partitioning, the analysis of dynamical processes based on eigenvectors and eigenvalues, or the computation of centrality measures that are based on eigenvalue problems.

Notably, the property of transitivity of paths in static networks does *not* extend to time-respecting paths in temporal networks. Here, two time-respecting paths $(v_0, v_1; t_1), \ldots, (v_{k-1}, v_k; t_k)$ and $(v_k, v_{k+1}; t_{k+1}), \ldots, (v_{l-1}, v_l; t_l)$ only translate into a time-respecting path between v_0 and v_l if $t_k < t_{k+1}$ and, assuming that we impose a maximum time difference δ, if $0 < t_{k+1} - t_k \leq \delta$.

[1] Note though that this transitive path may or may not be the shortest path between the two nodes.

The simple observation that transitivity of paths holds in static networks, while it does not necessarily hold in temporal networks implies that by an analysis of static, time-aggregated networks, we may overestimate transitivity in temporal networks. We can again illustrate this using our simple example of Figure 1, which shows two temporal networks G_1 and G_2 that are both consistent with the same (weighted) time-aggregated network shown in Figure 1c. Here, judging from the presence of a path $(a, c), (c, d)$ in the time-aggregated network, we may think that a time-respecting path connecting node a to d exists in the underlying temporal network. Looking at the two temporal networks G_1 and G_2 shown in Figures 1a and 1b respectively, we see that at least for small values for the maximum time difference δ (such as $\delta = 1$) a corresponding time-respecting path only exists in the temporal network G_1, while it is absent in G_2.

3 Higher-order aggregate networks

In the previous section we have seen that for large maximum time differences δ we expect the shortest time-respecting paths to be rather similar to the shortest paths in a static, time-aggregated representation. This is an intuitive result since, by using large maximum time differences δ, we apply an implicit "aggregation" of time stamps which may nevertheless be far apart in the temporal dimension. At the same time, we observe that for small values of δ the temporal characteristics of the network result in time-respecting path structures that are markedly different from those in the static, time-aggregated network. As argued above, this implies that dynamical processes which evolve at time scales similar to that of the temporal network will be significantly affected by these path structures. It further questions the usefulness of path-based centrality measures that are computed based on the commonly used time-aggregated representation of temporal networks.

In this section, we introduce *higher-order time-aggregated networks*, a simple yet powerful abstraction of temporal networks which can be used to address some of the aforementioned problems. It can be seen as a simple generalization of the usual *first-order* time-aggregated representation introduced in Section 2, and it has recently been shown to provide interesting insights about the evolution of dynamical processes in temporal networks [17].

3.1 kth order aggregate networks

The key idea behind this abstraction is that the commonly used time-aggregated network is the simplest possible time-aggregated representation whose weighted links capture the frequencies of time-stamped links. Considering that each time-stamped link is a time-respecting path of length one, it is easy to generalize this abstraction to *higher-order time-aggregate networks* in which weighted links capture the frequencies of longer time-respecting paths. For a temporal network $G^T = (V, E^T)$ we thus formally define a kth order time-aggregated (or simply aggregate) network as a tuple $G^{(k)} = (V^{(k)}, E^{(k)})$ where $V^{(k)} \subseteq V^k$ is a set of node k-tuples and $E^{(k)} \subseteq V^{(k)} \times V^{(k)}$ is a set of links. For simplicity, we call each of the k-tuples $v = v_1 - v_2 - \ldots - v_k$ $(v \in V^{(k)}, v_i \in V)$ a *kth order node*, while each link $e \in E^{(k)}$ is called a *kth order link*. We further assume that a kth order link (v, w) between two kth order nodes $v = v_1 - v_2 - \ldots - v_k$ and $w = w_1 - w_2 - \ldots - w_k$ exists if they overlap in exactly $k - 1$ elements such that $v_{i+1} = w_i$ for $i = 1, \ldots, k - 1$. Resembling so-called *De Bruijn graphs* [23], the basic idea behind this construction is that each kth order link (v, w) represents a possible time-respecting path of length k in the underlying temporal network, which connects node v_1 to node w_k via k time-stamped links

$$(v_1, v_2 = w_1; t_1), \ldots, (v_k = w_{k-1}, w_k; t_k). \quad (1)$$

In analogy to the weights in a usual (first-order) aggregate representation, we further define the weights of such kth order links by the frequency of the underlying time-respecting paths in the temporal network. Considering a maximum time difference δ and two kth order nodes $v = v_1 - v_2 - \ldots - v_k$ and $w = w_1 - w_2 - \ldots - w_k$ we thus define

$$\omega(v, w) := |P(v, w, \delta)|$$

where

$$P = \{(v_1, v_2 = w_1; t_1), \ldots, (v_k = w_{k-1}, w_k; t_k) : 0 < t_{i+1} - t_i \leq \delta\},$$

is the set of all time-respecting paths in the temporal network that (i) consist of the sequence of links indicated in equation (1), and (ii) are consistent with a given maximum time difference of δ.

The higher-order aggregate network construction introduced above has a number of advantages. First and foremost, it provides a simple static abstraction of a temporal network which can be studied by means of standard methods from (static) network analysis. Each static path of length l in a kth order aggregate network can be mapped to a time-respecting path of length $k+l-1$ in the original network. Importantly, and different from a first-order representation, kth order aggregate networks allow to capture *non-Markovian characteristics* of temporal networks. In particular, they allow to represent temporal networks in which the kth time-stamped link $(v_k = w_{k-1}, w_k)$ on a time-respecting path depends on the $k - 1$ previous time-stamped links on this path. With this, we obtain a simple static network topology that contains information both on the presence of time-stamped links in the underlying temporal network, as well as on the *ordering* in which sequences of k of these time-stamped links occur.

3.2 Example: second-order aggregate networks

In the following, we illustrate our approach by constructing second-order aggregate representations of the two temporal networks G_1 and G_2 shown in Figure 1. Both G_1

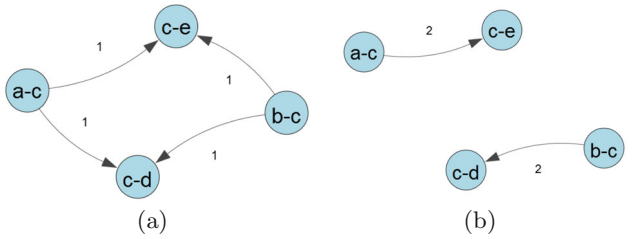

Fig. 2. Second-order aggregate networks $G^{(2)}$ corresponding to the two temporal networks shown in Figure 1. (a) Temporal network G_1. (b) Temporal network G_2.

and G_2 are consistent with the same first-order time-aggregated network. We can easily generate second-order time-aggregated networks of the two temporal networks by extracting all time-respecting paths of length two (and assuming a given maximum time difference δ). For simplicity, in the following we limit our study to $\delta = 1$. For the temporal network G_1 shown in Figure 1a, we observe the following four different time-respecting paths of length two:

$$(a, c; 1),\ (c, e; 2)$$
$$(b, c; 4),\ (c, d; 5)$$
$$(b, c; 7),\ (c, e; 8)$$
$$(a, c; 10),\ (c, d; 11).$$

Based on the definition of links and link weights outlined above, we thus obtain the following four weighted second-order links:

$$\omega(a - c, c - e) = 1$$
$$\omega(b - c, c - d) = 1$$
$$\omega(b - c, c - e) = 1$$
$$\omega(a - c, c - d) = 1.$$

The resulting second-order network is depicted in Figure 2a Applying the same methodology to the temporal network G_2 shown in Figure 1b we obtain the following four time-respecting paths of length two

$$(a, c; 1),\ (c, e; 2)$$
$$(b, c; 4),\ (c, d; 5)$$
$$(b, c; 7),\ (c, d; 8)$$
$$(a, c; 10),\ (c, e; 11),$$

from which we obtain the following two weighted second-order links:

$$\omega(a - c, c - e) = 2$$
$$\omega(b - c, c - d) = 2.$$

The resulting second-order aggregate network is shown in Figure 2b. Here we observe that, even though the two temporal networks G_1 and G_2 only differ in the order of two time-stamped links, the resulting second-order aggregate network is markedly different. The second-order network of G_1 indicates time-respecting paths connecting node a to both nodes e and d (both paths passing via node c). In particular, this corresponds to the connectivity that we would expect based on the transitivity of static paths in the first-order aggregate network shown in Figure 1c. The second-order network shown in Figure 2b reveals that the transitive path $(a, c), (c, d)$ in the first-order aggregate network does not translate to a time-respecting path in the temporal network G_2.

Clearly, the second-order aggregate networks illustrated above are only a special, particularly simple type of general, higher-order aggregate networks. Nevertheless, in the following section we will demonstrate that it contains important information about the causal topology of temporal networks which can help us in the analysis of temporal networks.

In what follows, we will thus provide an in-depth study of second-order aggregate representations of six empirical data sets that will be introduced in the following section. Here, we will particularly focus on the question how second-order aggregate networks can foster the calculation of approximate measures for path-based node centralities in temporal networks.

4 Temporal node centralities in second-order aggregate networks

Having introduced the abstraction of higher-order aggregate networks in Section 3, let us now demonstrate the use of a *second-order aggregate representation* for the study of path-based centralities in temporal networks. We will study this question using the following six, publicly available empirical data sets representing different types of temporal networks: (AN) covers time-stamped antenna-antenna interactions inferred from a filming of ants in an ant colony [24]; (EM) represents time-stamped E-Mail exchanges between employees in a manufacturing company [25]; (HO) covers time-stamped proximity interactions between patients and medical staff in a hospital [26]; (RM) is based on time-stamped social interactions between students and academic staff at a university campus [27]; (LT) has been reconstructed from data on passenger itineraries in the London Tube metro system available through the Rolling Origin and Destination Survey of the Transport of London [28], and (FL) was constructed based from data on flight itineraries of passengers on domestic flights in the United States available from the Bureau of Transportation Statistics [29]. A detailed description about the processing of these data sets and the extraction of time-stamped network data is available in reference [17], which is why we omit an elaborate discussion here.

Regarding the choice of a reasonable maximum time difference δ for the notion of shortest time-respecting paths as discussed in Section 2, we emphasize that the choice of this parameter needs to be adapted to the inherent time scale of the network evolution in each of the six data sets individually. In general, such a choice is non-trivial as it heavily influences (i) whether or not pairs

of nodes can reach each other, and (ii) to what extent temporal characteristics influence the structures of time-respecting paths. In particular, for too small choices of δ the definition of time-respecting paths is likely to be too restrictive and almost no paths will be found [1,19]. Contrariwise, the choice of a too large value for δ results in the fact that we effectively "aggregate" the time-stamped sequence of links, thus discarding information about the detailed ordering and timing of links. For our analysis, for each of the six data sets individually, we have thus chosen the minimum parameter δ for which we still obtain a topology of time-respecting paths that is strongly connected, thus ensuring that we can compute reasonable measures of path-based centralities while retaining as much of the temporal characteristics as possible (c.f. details in Ref. [17]).

In the remainder of this section, we will focus our analysis on three widely adopted path-based notions of centrality, namely (i) betweenness, (ii) closeness and (iii) reach centrality. The rationale behind this choice is that all of these three measures can easily be computed based on paths in time-aggregated networks, while they additionally facilitate a straight-forward extension to temporal networks based on the notion of shortest time-respecting paths (c.f. similar extensions studied in Refs. [1,20,21]). In the following, we first formally define the *temporal* betweenness, closeness and reach centrality of nodes. We then compute the resulting measures for all nodes based on the actual shortest time-respecting paths in the time-stamped link sequences in our six data sets (and using the individually determined maximum time difference δ). The resulting centrality scores are considered as the *ground-truth* against which we then compare the centrality scores resulting from the application of the same centrality measures to (i) the commonly used (first-order) time-aggregated representation, and (ii) a second-order aggregate network representation of the corresponding temporal network.

4.1 Temporal betweenness centrality

We first address the question to what extent the temporal betweenness centrality of nodes in a temporal network can be approximated by means of static betweenness centralities calculated based on static, time-aggregated representations. To this end, we first formally define the temporal betweenness centrality of a node in a temporal network. According to the common definition, the (unnormalized) betweenness centrality of a node v is simply calculated as the total number of shortest paths passing through node v [30]. Highlighting the fact that we can directly apply this measure to first-order time-aggregated networks, we thus define the *first-order betweenness centrality* $\mathrm{BC}^{(1)}(v)$ of a node v as:

$$\mathrm{BC}^{(1)}(v) := \sum_{u \neq v \neq w} |P^{(1)}(u,w;v)|, \tag{2}$$

where $P^{(1)}(u,w;v)$ denotes the set of those shortest paths from node u to w in a static network that pass through node v.

Applying this idea to temporal networks, a straightforward way to define the *temporal betweenness centrality* of a node is to count all shortest *time-respecting paths* passing through it. However, and as mentioned in Section 2, temporal networks introduce the complication that, in order to unambiguously define shortest time-respecting paths, we need to include a start time t_0 starting from which time-respecting paths are to be considered. For each pair of nodes u, v and each start time t_0 we can thus directly define an instantaneous distance function for a temporal network as:

$$\mathrm{dist}^{\mathrm{temp}}(u,v,t_0) := \mathrm{len}(p), p \in P^{\mathrm{temp}}(u,v,t_0), \tag{3}$$

where $P^{\mathrm{temp}}(u,v,t_0)$ denotes the set of shortest time-respecting paths from u to v that start at time t_0 (and which are consistent with a given maximum time difference δ). Based on this instantaneous definition of shortest time-respecting paths, we can further define a distance function that gives the minimum distance across *any* start time as follows:

$$\mathrm{dist}^{\mathrm{temp}}(u,v) := \min_{t_0} \mathrm{dist}^{\mathrm{temp}}(u,v,t_0). \tag{4}$$

With this we can further define the set of shortest time-respecting paths across all start times as:

$$\begin{aligned} P^{\mathrm{temp}}(u,v) : &= \bigcup_{t_0} \{p \in P^{\mathrm{temp}}(u,v,t_0) | \mathrm{len}(p) \\ &= \mathrm{dist}^{\mathrm{temp}}(u,v)\}, \end{aligned} \tag{5}$$

i.e. we only consider those (instantaneous) shortest time-respecting paths whose lengths correspond to the minimum shortest time-respecting length across *all* possible start times. We can now define the *temporal betweenness centrality* $\mathrm{BC}^{\mathrm{temp}}(v)$ of a node v in analogy to equation (2) as:

$$\mathrm{BC}^{\mathrm{temp}}(v) := \sum_{u \neq v \neq w} |P^{\mathrm{temp}}(u,w;v)|, \tag{6}$$

where $P^{\mathrm{temp}}(u,w;v)$ denotes the set of those shortest time-respecting paths across all start times which connect node u to w and which pass through node v.

Let us illustrate this definition using the temporal networks shown in Figures 1a and 1b. Applying the static betweenness centrality as defined in equation (2) to the first-order aggregate network shown in Figure 1c, we find that for node c we have $\mathrm{BC}^{(1)}(c) = 4$, while for all other nodes we have a betweenness centrality of zero. Again assuming $\delta = 1$, for the temporal betweenness centrality of node c in network G_1 shown in Figure 1a, we find that indeed four shortest time-respecting paths pass through node c, i.e. we have $\mathrm{BC}^{\mathrm{temp}}(c) = 4$ while we again have a zero temporal betweenness centrality for all other nodes. Notably, in this particular case the temporal betweenness

Table 1. Pearson and Kendall-Tau rank correlation coefficients between temporal betweenness centrality (ground truth) and betweenness centrality calculated based on the first-order aggregate network and the second-order aggregate network. Values in parentheses indicate the p-value.

	$BC^{temp} \sim BC^{(1)}$		$BC^{temp} \sim BC^{(2)}$	
	Pearson	Kendall-Tau	Pearson	Kendall-Tau
E-Mail (EM)	0.80 (3.29e-22)	0.73 (8.36e-26)	0.97 (7.52e-60)	0.74 (1.11e-26)
Ants (AN)	0.82 (3.49e-16)	0.64 (2.05e-13)	0.80 (1.96e-14)	0.59 (1.94e-11)
Hospital (HO)	0.93 (2.39e-23)	0.81 (1.18e-17)	0.96 (2.36e-30)	0.87 (5.55e-20)
RealityMining (RM)	0.95 (2.83e-30)	0.62 (7.28e-12)	0.93 (3.74e-26)	0.75 (1.12e-16)
London Tube (LT)	0.85 (2.58e-37)	0.66 (1.22e-29)	0.87 (3.28e-42)	0.71 (9.32e-34)
Flights (FL)	0.99 (6.91e-108)	0.66 (9.09e-26)	0.99 (2.66e-98)	0.65 (4.25e-25)

centralities of nodes correspond to the betweenness centralities of nodes calculated based on the first-order time-aggregated network. This happens because all paths in the first-order aggregate network have a counterpart in terms of a shortest time-respecting path.

However, in Section 2 we have seen that, in general, shortest time-respecting paths in temporal networks may not coincide with shortest paths in the (first-order) time-aggregated network. As a consequence, the temporal betweenness centralities of nodes may differ from the first-order betweenness centralities calculated from a static, first-order aggregate representation. This can be seen for the temporal network G_2 shown in Figure 1b. Based on the temporal sequence of time-stamped links, here we find only two different shortest time-respecting paths passing through node c, namely one connecting node a via c to e and a second one connecting node b via c to d. The two additional shortest time-respecting paths found in G_1 are absent in G_2, therefore in G_2 node c has a temporal betweenness centrality $BC^{temp}(c) = 2$, thus being, at least from the perspective of temporal betweenness centrality, less important than in G_1.

In the following we study the question to what extent first-order betweenness centralities can be used as a proxy for the temporal betweenness centralities of nodes in our six data sets of real-world temporal networks. In particular, we study this question in the following way: For each node v in the six data sets we calculate (i) the first-order betweenness centrality $BC^{(1)}(v)$ based on the first-order aggregate network, as well as (ii) the (ground truth) temporal betweenness centrality $BC^{temp}(v)$ based on actual shortest time-respecting paths in the temporal network. We then assess the correlation between both measures by computing the Pearson correlation coefficient (as well as the corresponding p-value) for the sequence of paired values $(BC^{(1)}(i), BC^{temp}(i))$ for all nodes $i \in V$.

Since centrality scores of nodes in networks are often used and interpreted in a relative fashion, we further perform an additional analysis that accounts for variations in the actual centrality values, which however may not affect the relative importance of nodes. For this, we first rank nodes according to their temporal and first-order betweenness centralities respectively. We then calculate the Kendall-Tau rank correlation coefficient in order to quantitatively assess to what extent nodes are ranked similarly according to both notions of centrality (even though the actual centrality values for these nodes may differ).

The results of this analysis are shown in the left column of Table 1, in which we report both the Pearson as well as the Kendall-Tau rank correlation coefficients between the temporal and the first-order betweenness centralities of nodes for each of the six data sets introduced above. Here, a first interesting result is that both the Pearson and the Kendall-Tau rank correlation coefficients exhibit a large variation between 0.80 and 0.99, as well as 0.62 and 0.81, respectively. The results indicate that, depending on the characteristics of the underlying temporal network, temporal betweenness centralities can be reasonably well approximated by first-order betweenness centrality for some data sets (e.g., for (FL), (HO), (RM)) while such an approximation should be taken with caution for other data sets.

Based on these results it is reasonable to ask if we can better approximate temporal centrality, especially for those data sets where the correlation between the first-order and the temporal betweenness centrality is comparably weak. In Section 3 we have argued that the generalization of higher-order aggregate networks allows to construct static representations of temporal networks that capture both temporal and topological characteristics that emerge from the ordering of links and the statistics of time-respecting paths. Focusing on a second-order representation, in the remainder of this section we will study to what extent second-order aggregate networks can be used in the analysis of temporal node centralities.

Importantly, such an analysis is facilitated by the fact that second-order aggregate networks are *static networks*, which allows for a straight-forward application of standard centrality measures to the second-order topology. In the case of second-order aggregate networks, applying standard centrality measures we obtain centrality values for higher-order nodes (v, w), each of the higher-order nodes being a k-tuple of nodes in the first-order network. In order to arrive at a centrality measure for the original (first-order) nodes, we thus must project this measure to the level of nodes in the first-order network.

Luckily, this can be done in a simple way which we outline in the following: For a second-order network $G^{(2)} = \left(V^{(2)}, E^{(2)}\right)$, let us first define a second-order distance function $dist^{(2)}(v, w)$ which, for each pair of *first-order*

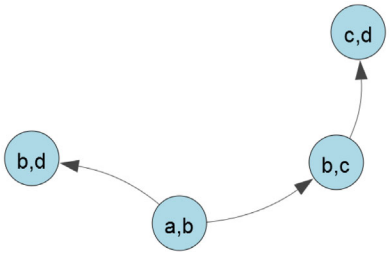

Fig. 3. Simple example for a second-order aggregate network.

nodes $v, w \in V^{(1)}$, gives the length of a shortest path based on the topology of the *second-order* aggregate network as:

$$\mathrm{dist}^{(2)}(v,w) := \min_{\substack{x,y \in V^{(2)} \\ x=v-* \\ y=*-w}} L^{(2)}(x,y) + 1, \tag{7}$$

where $L^{(2)}(x,y)$ denotes the length of a shortest path between the second-order nodes $x, y \in V^{(2)}$. The rationale behind this definition is that in the second-order aggregate network, we can have multiple shortest paths with different lengths between different second-order nodes, which nevertheless map to paths between a single pair of first-order nodes. As an example, consider the two first-order nodes a and d in the simple second-order network shown in Figure 3. Here we observe that, from the perspective of second-order nodes, both $(a-b, b-d)$ as well as $(a-b, b-c), (b-c, c-d)$ are shortest paths (between different pairs of nodes) in the second-order network with lengths $L^{(2)}(a-b, b-d) = 1$ and $L^{(2)}(a-b, c-d) = 2$ respectively. However, from the perspective of first-order nodes both of these second-order paths connect node a to node d (via paths of length 2 and 3, respectively). Using the definition from equation (7) thus allows us to correctly calculate the second-order distance between a and d as $\mathrm{dist}^{(2)}(a,d) = L^{(2)}(a-b, b-d) + 1 = 2$.

The above definition of a second-order distance function now allows us to define a *second-order betweenness centrality* $\mathrm{BC}^{(2)}(v)$ of a node v based on equation (2). For this, we simply count all second-order shortest paths between two nodes u and w which (i) pass through node v, and (ii) whose length corresponds to the second-order distance $\mathrm{dist}^{(2)}(u,v)$. Formally, we define

$$\begin{aligned}\mathrm{BC}^{(2)}(v) := & \sum_{\substack{x \neq y \in V \\ u-x \in V^{(2)} \\ y-w \in V^{(2)}}} |\{p \in P^{(2)}(u-x, y-w; v) : \mathrm{len}(p) \\ & = \mathrm{dist}^{(2)}(u,w)\}|, \end{aligned} \tag{8}$$

where, in analogy to $P^{(1)}(u,w;v)$ above,

$$P^{(2)}(u-x, y-w; v)$$

denotes the set of all shortest paths in the second-order network that connect node $u-x$ to $y-w$ and that pass through a first-order node v.

With this, we have defined a second-order betweenness centrality which allows to calculate node centralities in a way that incorporates the causal topology as captured by the second-order aggregate network. Let us again illustrate this approach using the simple examples shown in Figure 1. For the temporal network G_1 we can compute a second-order betweenness centrality based on the second-order network shown in Figure 2a. Here we observe a total of four shortest paths between pairs of nodes in the second-order network, namely:

$$\begin{aligned}&(a-c, c-e),\\ &(a-c, c-d),\\ &(b-c, c-d),\\ &(b-c, c-e).\end{aligned}$$

For each node in the first-order network, we can now count the number of second-order shortest paths that they are on, obtaining $B^{(2)}(c) = 4$ while $B^{(2)}(x) = 0$ for all nodes $x \neq c$. In this particular case, the second-order betweenness centrality values exactly correspond both to the temporal as well as the first-order betweenness centralities. Again, this is different for the temporal network G_2 shown in Figure 1b. Considering the second-order aggregate network shown in Figure 2b, we only find the following two shortest paths in the second-order aggregate network

$$\begin{aligned}&(b-c, c-d)\\ &(a-c, c-e)\end{aligned}$$

thus obtaining $\mathrm{BC}^{(2)}(c) = 2$. Here, we find that while the second-order betweenness centralities in G_2 corresponds to the temporal betweenness centralities, they differ from those calculated from the first-order aggregate network. The reason for this is that in the example G_2 shortest time-respecting paths of length two differ from what we would expect based on the first-order network.

We emphasize that the exact correspondence between the second-order and the temporal betweenness centralities in the examples discussed above is because we have no shortest time-respecting paths of length three or longer, whose presence could differ from what we expect based on the second-order network. To what extent this affects the applicability of second-order aggregate networks in real-world scenarios is not clear and thus requires a further investigation. In the following, we thus study to what extent second-order betweenness centrality can be used to approximate the temporal betweenness centralities of nodes in the six real-world data sets studied above. For this, we first construct a second-order aggregate network as introduced in Section 3. We then calculate the betweenness centrality values $\mathrm{BC}^{(2)}(v)$ of all nodes v as described above, comparing the resulting centralities with the (ground-truth) temporal betweenness centralities $\mathrm{BC}^{\mathrm{temp}}(v)$.

The results of this analysis are shown in the right column of Table 1. Here we find that for most of the data sets, second-order betweenness centralities are correlated with the true, temporal betweenness centralities in a stronger way than the corresponding first-order approximation of

betweenness centrality. For the (EM) data sets capturing E-Mail exchanges between employees in a manufacturing company, we observe an increase of the Pearson correlation coefficient ρ from 0.80 to 0.97, while the associated Kendall-Tau rank correlation coefficient τ increases rather mildly from 0.73 to 0.74. We attribute this to the fact that the second-order aggregate network better captures the structures of time-respecting paths in the temporal network compared to the first-order network. For the two data sets (HO) and (LT) we observe a similar increase both in the Pearson and the Kendall-Tau rank correlation coefficients, while the values remain largely unchanged for the (FL) data set. In particular, for the latter data set the first-order betweenness centrality already exhibits a correlation coefficient of 0.99 which indicates that in this particular case temporal characteristics do not significantly alter the structure of shortest time-respecting paths. For the two data sets (AN) and (RM) we observe a small decrease in the Pearson correlation values for the second-order approximation. Notably, for (RN) the decrease from 0.97 to 0.95 is accompanied by an increase of the Kendall-Tau coefficient from 0.62 to 0.75. This indicates that, even though the actual values of second-order betweenness centralities may be less correlated with temporal betweenness centralities than the first-order betweenness centralities, the second-order betweenness centralities provides us with a significantly better perspective on the *relative* importance of nodes.

Finally, for the (AN) data set we note that both the Pearson and the Kendall-Tau rank correlation coefficients are worse for the second-order betweenness centralities. While the interesting question in what respect the temporal characteristics of (AN) differ from those of the other temporal networks remains to be investigated in more detail, we expect this result to be related to non-stationary properties. We particularly observe that some of the nodes (i.e. ants) are only active during certain phases of the observation period. This imposes a natural ordering of interactions which particularly prevents nodes which are only active during an early phase to be reachable from nodes which are only active at a later phase.

4.2 Temporal closeness centrality

Let us now turn our attention to *closeness centrality*, which captures a node's average distance to all other nodes in a network. For a directed, static (first-order aggregate) network the closeness centrality of a node v is commonly defined as:

$$\mathrm{ClC}^{(1)}(v) = \sum_{u \neq v} \frac{1}{\mathrm{dist}^{(1)}(u,v)}, \tag{9}$$

where the distance function $\mathrm{dist}^{(1)}(u,v)$ denotes the distance, i.e. the length of a shortest path, from node u to v in the first-order aggregate network.

We can easily define a temporal version of closeness centrality based on the temporal distance function $\mathrm{dist}^{\mathrm{temp}}(u,v)$ which we have defined in equation (4) in the context of temporal betweenness centrality. Here, we remind the reader that the function $\mathrm{dist}^{\mathrm{temp}}(u,v)$ captures the minimum length of a shortest time-respecting path across all possible start times t_0. Using this temporal distance function, we can apply the standard definition in equation (9) and define the *temporal closeness centrality* of a node v in a temporal network as:

$$\mathrm{ClC}^{\mathrm{temp}}(v) = \sum_{u \neq v} \frac{1}{\mathrm{dist}^{\mathrm{temp}}(u,v)}. \tag{10}$$

Let us again illustrate this definition using the temporal networks shown in Figure 1. Node e in the temporal network G_1 shown in Figure 1a can be reached from nodes a and b via two shortest time-respecting paths of length two, as well as from node c via a shortest time-respecting path of length one. For the temporal closeness centrality, we thus find $\mathrm{ClC}^{\mathrm{temp}}(e) = 2$. It is easy to confirm that this corresponds to the first-order closeness centrality of node e. Again a mere reordering of links can change the closeness centralities of nodes, as can be seen in the temporal network G_2 shown in Figure 1b. Here, we see that node e can only be reached from node a via a shortest time-respecting path of length two, as well as from node c via a shortest time-respecting path of length one. For node e in the temporal network G_2 we thus find a temporal closeness centrality $\mathrm{ClC}^{\mathrm{temp}}(e) = 1.5$, highlighting that it is, at least from the perspective of closeness centrality, less "important" than in the temporal network G_1.

Considering the example above we see that, due to the ordering and timing of links, first-order closeness centralities can be a misleading proxy for the temporal closeness centralities of nodes in temporal networks. In the following we thus again empirically study this question using our six data sets on temporal networks. We again use the temporal closeness centralities $\mathrm{ClC}^{\mathrm{temp}}(v)$ of nodes as the ground truth, then studying whether temporal closeness centralities can reasonably be approximated by first-order closeness centralities $\mathrm{ClC}^{(1)}(v)$. The results of this analysis are shown in the left column of Table 2, which reports the observed Pearson and Kendall-Tau rank correlation coefficients for each of the six data sets.

We observe again that the answer to the question of how well temporal closeness centralities can be approximated by first-order static closeness centralities depends on the actual data set. The lowest Pearson correlation coefficient of 0.91 is obtained for the (FL) and the (AN) data sets, while the highest Pearson correlation coefficient of 0.98 is obtained for (LT). The lowest Kendall-Tau rank correlation coefficient is 0.75 for (AN), while the highest value of 0.87 is achieved for (LT). We further observe that, compared to betweenness centralities, we generally obtain conceivably larger correlation values between temporal and first-order closeness centralities. This can intuitively be explained by the fact that, while temporal betweenness centralities are influenced by the actual *structure* of shortest time-respecting paths, temporal closeness centralities are merely influenced by their lengths. We thus expect temporal closeness centrality to be insensitive to characteristics of temporal networks that change the structure

Table 2. Pearson and Kendall-Tau rank correlation coefficients between temporal closeness centrality (ground truth) and closeness centrality calculated based on the first-order aggregate network and the second order aggregate network. Values in parentheses indicate the p-value.

	$\mathrm{ClC}^{\mathrm{temp}} \sim \mathrm{ClC}^{(1)}$		$\mathrm{ClC}^{\mathrm{temp}} \sim \mathrm{ClC}^{(2)}$	
	Pearson	Kendall-Tau	Pearson	Kendall-Tau
E-Mail (EM)	0.93 (4.74e-44)	0.79 (4.96e-30)	0.98 (2.52e-71)	0.92 (1.54e-40)
Ants (AN)	0.91 (1.67e-24)	0.75 (1.54e-17)	0.96 (2.05e-35)	0.83 (4.80e-21)
Hospital (HO)	0.96 (2.09e-29)	0.83 (1.88e-18)	0.99 (1.46e-40)	0.90 (1.76e-21)
RealityMining (RM)	0.96 (1.03e-33)	0.77 (1.99e-17)	0.99 (1.64e-51)	0.89 (5.30e-17)
London Tube (LT)	0.98 (1.33e-91)	0.87 (2.57e-49)	0.98 (3.26e-92)	0.87 (1.07e-49)
Flights (FL)	0.91 (3.35e-46)	0.81 (1.88e-18)	0.97 (4.57e-75)	0.93 (9.57e-50)

of paths but not their lengths, hence explaining the larger correlation coefficients.

Let us now study whether we can better approximate temporal closeness centralities using a generalization which is calculated based on the static, second-order aggregate representation of a temporal network. For this we first introduce how closeness centralities of nodes can be calculated based on a second-order aggregate network. We recall that in equation (7) we have defined a second-order distance function $\mathrm{dist}^{(2)}(v,w)$ which provides us with the distance between (first-order) nodes based on shortest paths in a second-order aggregate network. This distance function allows us to directly define a *second-order closeness centrality* $\mathrm{ClC}^{(2)}(v)$ as:

$$\mathrm{ClC}^{(2)}(v) = \sum_{u \neq v} \frac{1}{\mathrm{dist}^{(2)}(u,v)}, \tag{11}$$

i.e. for each node v in a network, we simply sum the inverse of the distances to all nodes according to the topology of the second-order aggregate network.

Again, we illustrate the notion of second-order closeness centrality using the two illustrative examples of temporal networks shown in Figure1. Figure 2a shows the second-order aggregate network corresponding to the temporal network G_1 shown in Figure 1a. Here we find that the second-order node $c-e$ can be reached via two shortest paths

$$(b-c),(c-e)$$
$$(a-c),(c-e)$$

of length one from the second-order nodes $b-c$ and $a-c$. Furthermore, we have an additional second-order "path" of length zero from node $c-e$ to itself. Using the second-order distance function as defined in equation (7), we thus infer the following values:

$$\mathrm{dist}^{(2)}(b,e) = 2$$
$$\mathrm{dist}^{(2)}(a,e) = 2$$
$$\mathrm{dist}^{(2)}(c,e) = 1$$

from which we calculate the second-order closeness centrality of node e as $\mathrm{ClC}^{(2)}(c) = 2$.

Again, in this particular example the second-order closeness centrality corresponds both to the temporal and the first-order closeness centrality. This is different in the second-order network shown in Figure 2b, which corresponds to the temporal network G_2 shown in Figure 1b. Here, we find that the second-order node $c-e$ can only be reached via a single shortest path $(a-c),(c-e)$ as well as via an additional second-order "path" of length zero from $e-c$ to itself. From this, we can calculate the following second-order distances

$$\mathrm{dist}^{(2)}(a,e) = 2$$
$$\mathrm{dist}^{(2)}(c,e) = 1$$

and for the second-order closeness centrality of node e we thus obtain $\mathrm{ClC}^{(2)}(c) = 1.5$, which coincides with the temporal closeness of node e in the underlying temporal network G_2.

Using the the second-order closeness centrality introduced above, let us now study the correlations between the temporal and the second-order closeness centralities of nodes in our six data sets. The results of this analysis are shown in the right column of Table 2. For five of the six data sets we observe significantly larger correlation coefficients than those reported for the first-order closeness centrality in Table 2. The largest increase of the Pearson correlation coefficient from 0.91 to 0.97 is achieved for the (FL) data set, while we observe no improvement of the (already large) Pearson correlation coefficient of 0.98 for (LT). We further observe significant increases in the Kendall-Tau rank correlation coefficients for all of the studied data sets, except for (LT) for which it remains the same. For the ranking of nodes in (EM), we find that a ranking based on second-order closeness centralities increases the Kendall-Tau rank correlation with the ground truth temporal centralities from 0.79 to 0.92, thus better representing the relative importance of nodes in the temporal network.

4.3 Temporal reach centrality

Concluding this section we finally study *reach centrality*, another notion of path-based centrality that captures the number of nodes that can be reached from a node via paths up to given maximum length s [31]. For static networks, such as a first-order aggregate network, we define

the *first-order* reach centrality of a node v as:

$$\mathrm{CoC}^{(1)}(v,s) := \sum_{w \in V} \Theta(\mathrm{dist}^{(1)}(v,w) - s), \qquad (12)$$

where $\Theta(\cdot)$ is the Heaviside function, $\mathrm{dist}^{(1)}(v,u)$ is the length of a shortest path from node v to u in the static, first-order network, and s is a parameter specifying up to which length paths should be considered. Clearly, the reach centrality $\mathrm{CoC}^{(1)}(v, s=1)$ of a node v is equal to its out-degree while $\mathrm{CoC}^{(1)}(v, s=\infty)$ is equal to the subset of nodes to which v is connected via directed paths of any length.

A *temporal reach centrality* can again easily be defined based on the notion of shortest time-respecting paths, as well as the temporal distance function $\mathrm{dist}^{\mathrm{temp}}(v,w)$ defined in equation (4). Here, for a given maximum time difference δ and a given value s, we are interested in how many different nodes can be reached via shortest time-respecting paths which have at most length s. In analogy to equation (12), we can thus define the *temporal reach centrality* $\mathrm{CoC}^{\mathrm{temp}}(v)$ of a node v as:

$$\mathrm{CoC}^{\mathrm{temp}}(v,s) := \sum_{w \in V} \Theta(\mathrm{dist}^{\mathrm{temp}}(v,w) - s). \qquad (13)$$

We want to highlight that with this definition of reach centrality, we focus on the temporal-topological characteristics introduced by the ordering of links, which is why base our definition on the *shortest* rather than the *fastest* time-respecting paths.

It is finally easy to see that a *second-order reach centrality* can be defined in analogy to second-order closeness centrality. For this, all we have to do is to replace the distance function in equation (12) by our previously defined second-order distance function, thus obtaining the following definition:

$$\mathrm{CoC}^{(2)}(v,s) := \sum_{w \in V} \Theta(\mathrm{dist}^{(2)}(v,w) - s). \qquad (14)$$

Using a value of $s = 2$, we again exemplify these definitions using our two illustrative examples. Let us first calculate the first-order reach centrality of node a based on the first-order aggregate network shown in Figure 1c. Here we find that there are paths of at most length $s = 2$ from node a to the three nodes c, d and e, from which we conclude $\mathrm{CoC}^{(1)}(a, s=2) = 3$. For the temporal reach centrality of node a in the temporal network G_1 shown in Figure 1a, we observe that there are time-respecting paths of at most length $s = 2$ from node a to the three nodes c, e and d. We hence conclude $\mathrm{CoC}^{\mathrm{temp}}(a, s=2) = 3$, finding that for G_1 the temporal reach centrality again corresponds to the first-order reach centrality. Again, this is not the case for the temporal network G_2 shown in Figure 1b. Here, node a is only connected to the nodes c and e via time-respecting paths of up to length two, which means that we have $\mathrm{CoC}^{\mathrm{temp}}(a, s=2) = 2$.

For the second-order reach centrality of node a in the temporal network G_1 let us now consider the second-order aggregate network shown in Figure 2a. Based on the shortest paths in the second-order network, we first find that the node $a-c$ is connected to two nodes $c-d$ and $c-e$ via shortest paths of length one. Furthermore, we find an additional shortest path of length zero which connects the second-order node $a-c$ to itself. Again, using our second-order distance function $\mathrm{dist}^{(2)}$ here we find the distances

$$\mathrm{dist}^{(2)}(a,c) = 1$$
$$\mathrm{dist}^{(2)}(a,e) = 2$$
$$\mathrm{dist}^{(2)}(a,d) = 2,$$

from which we conclude that three nodes c, e and d can be reached via paths of length at most two. From this we calculate the second-order reach centrality of node a in G_1 as $\mathrm{CoC}^{(2)}(a, s=2) = 3$. Applying the same arguments to the example network G_2 and the corresponding second-order aggregate network shown in Figure 2b, for the same three nodes we find the following second-order distances:

$$\mathrm{dist}^{(2)}(a,c) = 1$$
$$\mathrm{dist}^{(2)}(a,e) = 2$$
$$\mathrm{dist}^{(2)}(a,d) = \infty.$$

We thus obtain a second-order reach centrality of $\mathrm{CoC}^{(2)}(a, s=2) = 2$ which corresponds to the temporal reach centrality of node a in G_2.

In the following, we use the temporal reach centrality defined above as ground truth, while studying how well it can be approximated by first-order and second-order reach centralities calculated from the first- and second-order time-aggregated networks respectively. Different from the analyses for betweenness and closeness centralities, here we must additionally account for the fact that the reach centrality can be calculated for different values of the maximum path length s. This implies that the Pearson correlation coefficient ρ and the Kendall-Tau rank correlation coefficient τ must be calculated for each value of s individually. The results of this analysis are shown in Figure 4, which shows the obtained values for ρ and τ for the correlations between (i) the temporal and the first-order reach centralities (black lines), and (ii) the temporal and the second-order reach centralities (cyan lines) for each of the six data sets introduced above. Thanks to our choice of the maximum time difference δ, for all of our data sets both the underlying first- and second-order networks are strongly connected. Assuming that D is the diameter of the corresponding aggregate network, for all $s \geq D$ we thus necessarily arrive at a situation where the reach centralities of all nodes are identical. For the results in Figure 4 this implies that for any $s > D$ the correlation values are undefined since the first- (or second-)order centralities of all nodes are the same. We thus only plot the correlation coefficients τ and ρ for $s < D$, in which case they are well-defined.

For $s = 1$, the only time-respecting paths considered consist of single links, and thus the temporal reach centralities by definition exactly correspond to the reach

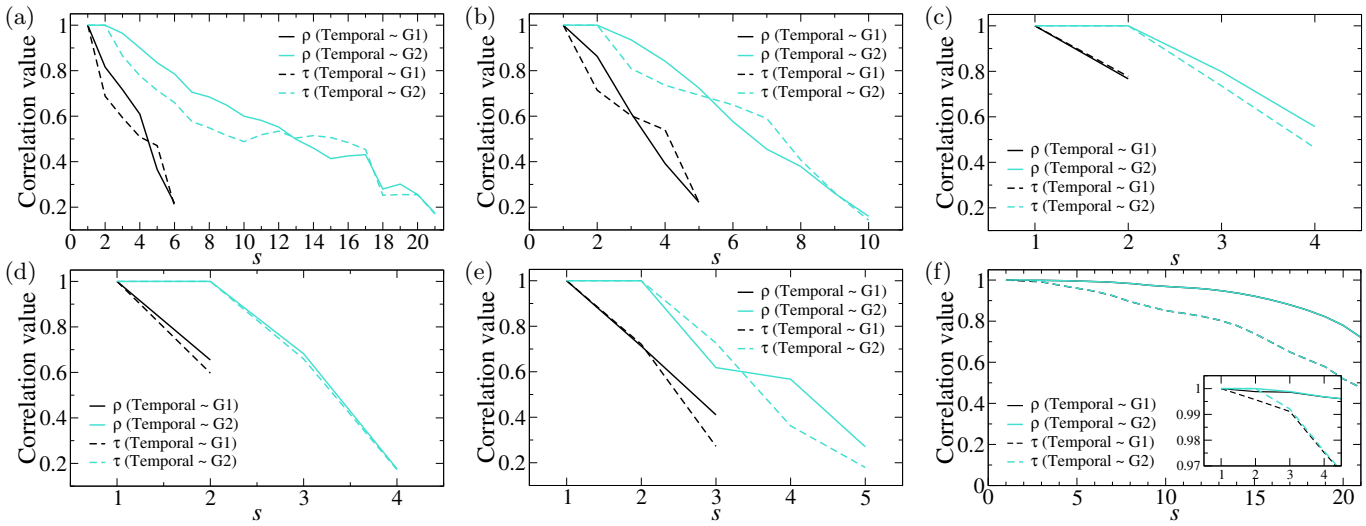

Fig. 4. Pearson ρ and Kendall τ correlation coefficients between the temporal and the first-order reach centralities (black lines) and the temporal and the second-order reach centralities (cyan lines) for (a) the Ants data set (AN), (b) the E-Mail data set (EM), (c) the Hospital data set (HO), (d) the Reality Mining dataset (RM), (e) the Flights data set (FL), and (f) the London Tube data set (LT). Inset: zoom to the area where there is a small deviation between values for the case of the London Tube data set.

centralities calculated from the first- and second-order topologies. Consequently, for $s = 1$ we have $\tau = 1$ and $\rho = 1$ both for the first- and the second-order reach centrality. For $s = 2$ there is, again by definition, no difference between the temporal and the second-order reach centralities however the correlation values for the first-order reach centrality decreases since the first-order aggregate network does not accurately represent the structure of time-respecting paths of length two. For values $s > 2$, ρ and τ decrease both for the first and the second-order centralities since neither representation can accurately represent time-respecting paths with lengths $s > 2$. However the results also highlight the important fact that second-order reach centralities better approximate temporal reach centralities for all values of $s > 2$.

We conclude this section by providing detailed results for the specific value of $s = 3$. The choice of a parameter $s > 2$ means that for the second-order reach centrality we will not trivially obtain correlation values of 1 because we would only consider time-respecting paths of length two which are captured in the second-order aggregate network. However, since the diameter of the first-order aggregate network for two of our systems (RM and HO) is equal to three, we can only report results on the correlations between the temporal and the first-order reach centralities for four data sets. The results for the first-order reach centrality with $s = 3$ are shown in in Table 3.

Remarkably, for the (LT) data sets we observe a perfect correlation with the temporal reach centrality, which means that for this data set reach centralities are seemingly not affected by the temporal characteristics of the system. This is different for (FL), for which we observe a small Pearson correlation of $\rho = 0.41$, with an associated $\tau = 0.27$. These results show that, for the (FL) data set, temporal characteristics of the data do not allow temporal reach centralities to be approximated based on the first-order aggregate network. For the second-order reach centralities shown in the right columns of Table 3, we observe a significant increase in both the Pearson and the Kendall-Tau correlation coefficients for all of the data sets, except for (LT). The largest increase of the Pearson correlation coefficient is again obtained for (EM), increasing from 0.61 to 0.94 with an associated increase of the Kendall-Tau correlation coefficient from 0.60 to 0.81. We thus conclude that again, second-order reach centralities better capture the true (temporal) importance of nodes than a simple first-order approximation.

5 Conclusion

In this article, we introduce a framework for the analysis of path-based notions of node centrality in temporal networks. We consider temporal versions of three path-based centrality measures which highlight the influence of the temporal-topological dimension introduced by the specific timing and ordering of time-stamped links in temporal networks. Using six data sets on real-world temporal networks, we study to what extent static notions of betweenness, closeness and reach centrality differ from their temporal counterparts. While for some data sets node centralities in the (first-order) time-aggregated, static network can be used as reasonable proxies for temporal centralities, our results show that for other data sets this is not the case. Here we find that an analysis of time-aggregated static networks that neglects the time dimensions yields misleading results about the importance of nodes.

In order to overcome these limitations, we utilize higher-order aggregate networks, a simple yet powerful

Table 3. Pearson and Kendall-Tau rank correlation coefficients between temporal reach centrality (ground truth) and reach centrality for $s = 3$ calculated based on the first-order aggregate network and the second-order aggregate network. Values in parentheses indicate the p-value.

	$\mathrm{CoC}^{\mathrm{temp}} \sim \mathrm{CoC}^{(1)}$		$\mathrm{CoC}^{\mathrm{temp}} \sim \mathrm{CoC}^{(2)}$	
	Pearson	Kendall-Tau	Pearson	Kendall-Tau
London Tube (LT)	1.00 (4.65e-168)	1.00 (9.32e-64)	1.00 (1.92e-173)	1.00 (7.00e-64)
Ants (AN)	0.72 (8.23e-11)	0.59 (1.38e-11)	0.96 (9.50e-36)	0.86 (6.40e-23)
E-Mail (EM)	0.61 (3.17e-11)	0.60 (3.55e-18)	0.94 (2.74e-44)	0.81 (1.52e-31)
RealityMining (RM)	NA	NA	0.68 (3.76e-09)	0.66 (2.78e-13)
Hospital (HO)	NA	NA	0.80 (7.95e-13)	0.74 (7.23e-15)
Flights (FL)	0.41 (4.68e-06)	0.27 (1.46e-05)	0.62 (1.44e-13)	0.73 (6.53e-31)

generalization of the commonly used time-aggregated static perspective on time-stamped network data which has originally been introduced to study dynamical processes in reference [17]. The basic idea of this construction is that a kth order aggregate network captures the statistics of time-respecting paths of length k, thus facilitating a higher-order analysis that incorporates both the topology and the ordering of links in temporal networks. We demonstrate the power of this framework through the definition of *second-order* centralities which can easily be calculated based on shortest paths in a second-order aggregate network. Despite the fact that these centralities can easily be calculated based on a simple static network structure, we find that the resulting second-order centrality measures better capture the true temporal centralities of nodes in the underlying temporal networks.

Closing this article, we finally highlight a number of open issues which we plan to consider in future work.

First and foremost, all of our results have been obtained based on simple *unweighted* notions of centralities, even though both the first-and second-order aggregate networks considered in our work naturally provide *weighted* links. Insofar, our results are based on an approach that does not incorporate full information about path statistics which is preserved by our higher-order aggregate network abstraction. We thus expect a future extension to weighted higher-order aggregate networks to capture the true temporal centralities of nodes even more closely.

Moreover, while we can in principle define higher-order networks of any order k, in our work we have focused on second-order representations and the corresponding generalizations of path-based centralities. The choice to limit our study to $k = 2$ is mainly due to available data which, for the six temporal networks studied in this work, are not guaranteed to provide meaningful statistics for time-respecting paths with larger lengths k that are the basis for a kth order aggregate network. Under what conditions higher-order aggregate networks with orders of $k > 2$ can help us to obtain even better approximations for temporal centralities is thus an open question that we will study in the future. Despite these open issues, we consider the fact that the simple second-order centrality measures introduced in our work already yield good approximations of the underlying temporal centralities a promising aspect of our framework.

As argued in Section 3, the construction of higher-order aggregate networks crucially depends on the notion of *time-respecting paths*. As such, a sensible choice of the maximum time difference δ – i.e., the maximum difference allowed between the time stamps of any two consecutive links that form a time-respecting path – is of particular importance. Depending on the temporal distribution of time-stamped links, choosing a δ that is too small may result in a negligible number of time-respecting paths, thus rendering the system in question (temporally) disconnected. Similarly, the choice of too large values for δ effectively leads to time-respecting paths that are independent of the precise timing and ordering of time-stamped links, thus effectively coarse-graining time and discarding temporal information. In general, the optimum choice of δ depends on (i) temporal characteristics of the temporal network under investigation, and (ii) the time scale of the dynamical processes one is interested in. In this study, we have used a rather simple and heuristic method, defining δ as the smallest value that still renders the system strongly connected in a temporal sense. In future work, a more principled approach which, e.g., inherently couples the choice of δ to a characterization of inter-event time statistics would be desirable.

Finally, an important general question that arises in the analysis of time-stamped network data is under which conditions a time-aggregated analysis is sufficient, as opposed to a detailed analysis of time-stamped links and time-respecting paths. Thanks to their simplicity, computational efficiency, and the availability of software tools, time-aggregated analyses of *static node centralities* are popular and widely used throughout different disciplines. However, the results of our analysis, as well as of similar studies on dynamical processes, community structures and node centralities [14,15,17,18] show that order correlations in real-world temporal networks crucially influence causality, thus potentially rendering such static analyses invalid. Through a calculation of *temporal node centralities* these temporal correlations can be included in the analysis of time-stamped data. However, especially for larger values of the maximum time difference δ, the extraction of all *shortest time-respecting* paths imposes computational costs that can be prohibitive for large data sets.

A particular benefit of our approach is that the calculation of second-order centrality measures is computationally efficient as it merely requires (i) the extraction of time-respecting paths of length two in the time-stamped data, and (ii) the calculation of shortest paths in a *static* second-order network. Therefore, we argue that our approach is a simple and efficient (static) approximation of temporal centralities which, compared to a calculation of *first-order* centralities, nevertheless provides significant additional insights into the temporal dimension of complex systems.

In summary, our results show that higher-order aggregate networks constitute a powerful abstraction for the study of temporal networks. Apart from an analysis of path structures and node centralities, as well as analytical studies of dynamical processes considered in reference [17], we argue that this framework provides broad perspectives for the development of novel, computationally efficient data mining techniques for time-stamped relational data.

IS acknowledges funding from the Swiss National Science Foundation (SNF), Grant No. CR31I1_140644/1, the Swiss State Secretariat for Education, Research and Innovation (SBFI), Grant No. C14.0036, as well as from the MTEC Foundation in context of the project "The Influence of Interaction Patterns on Success in Socio-Technical Systems". IS further acknowledges support from the European Cooperation in Science and Technology (COST) project TD1210. AG and NW acknowledge support from the FET project MULTIPLEX, Grant No. 317532.

References

1. P. Holme, J. Saramäki, Phys. Rep. **519**, 97 (2012)
2. P. Holme, Eur. Phys. J. B **88**, 234 (2015)
3. J.L. Iribarren, E. Moro, Phys. Rev. Lett. **103**, 038702 (2009)
4. M. Karsai, M. Kivelä, R.K. Pan, K. Kaski, J. Kertész, A.L. Barabási, J. Saramäki, Phys. Rev. E **83**, 025102 (2011)
5. L.E.C. Rocha, F. Liljeros, P. Holme, PLoS Comput. Biol. **7**, e1001109 (2011)
6. M. Starnini, A. Baronchelli, A. Barrat, R. Pastor-Satorras, Phys. Rev. E **85**, 056115 (2012)
7. N. Perra, A. Baronchelli, D. Mocanu, B. Goncalves, R. Pastor-Satorras, A. Vespignani, Phys. Rev. Lett. **109**, 238701 (2012)
8. N. Perra, B. Gonçalves, R. Pastor-Satorras, A. Vespignani, Sci. Rep. **2**, 469 (2012)
9. T. Hoffmann, M.A. Porter, R. Lambiotte, Random Walks on stochastic temporal networks, in *Temporal Networks, Understanding Complex Systems*, edited by P. Holme, J. Saramki (Springer, Berlin, Heidelberg, 2013), pp. 295−313
10. T. Takaguchi, N. Masuda, P. Holme, PLoS ONE **8**, e68629 (2013)
11. L.E.C. Rocha, V.D. Blondel, PLoS Comput. Biol. **9**, e1002974 (2013)
12. M. Karsai, N. Perra, A. Vespignani, Sci. Rep. **4**, 4001 (2014)
13. H.H. Jo, J.I. Perotti, K. Kaski, J. Kertész, Phys. Rev. X **4**, 011041 (2014)
14. H.H.K. Lentz, T. Selhorst, I.M. Sokolov, Phys. Rev. Lett. **110**, 118701 (2013)
15. R. Pfitzner, I. Scholtes, A. Garas, C.J. Tessone, F. Schweitzer, Phys. Rev. Lett. **110**, 198701 (2013)
16. R. Lambiotte, V. Salnikov, M. Rosvall, J. Complex Networks **3**, 177 (2015)
17. I. Scholtes, N. Wider, R. Pfitzner, A. Garas, C.J. Tessone, F. Schweitzer, Nat. Commun. **5**, 5024 (2014)
18. M. Rosvall, A.V. Esquivel, A. Lancichinetti, J.D. West, R. Lambiotte, Nat. Commun. **5**, 4630 (2014)
19. R. Pan, J. Saramäki, Phys. Rev. E **84**, 1 (2011)
20. H. Kim, R. Anderson, Phys. Rev. E **85**, 1 (2012)
21. T. Takaguchi, Y. Yano, Y. Yoshida, Eur. Phys. J. B **89**, 35 (2015)
22. D. Kempe, J. Kleinberg, A. Kumar, in *Proceedings of the thirty-second annual ACM symposium on Theory of computing ACM, 2000*, pp. 504−513
23. N.G. de Bruijn, Koninklijke Nederlandse Akademie v. Wetenschappen **49**, 758764 (1946)
24. B. Blonder, A. Dornhaus, PLoS ONE **6**, e20298 (2011)
25. R. Michalski, S. Palus, P. Kazienko, in *Business Information Systems*, Lecture notes in business information processing, edited by W. Abramowicz (Springer, Berlin, Heidelberg, 2011), Vol. 87, pp. 197–206
26. P. Vanhems, A. Barrat, C. Cattuto, J.F. Pinton, N. Khanafer, C. Regis, B.A. Kim, B. Comte, N. Voirin, PLoS ONE **8**, e73970 (2013)
27. N. Eagle, A. (Sandy) Pentland, Personal Ubiquitous Comput. **10**, 255 (2006)
28. Transport for London, Rolling Origin and Destination Survey (RODS) database (2014)
29. RITA TransStat Origin and Destination Survey database, available online 2014
30. L.C. Freeman, Sociometry **40**, 35 (1977)
31. S. Borgatti, M. Everett, J. Johnson, *Analyzing Social Networks* (SAGE Publications, 2013)