Applied Network Science

RESEARCH Open Access

Analytics for directed contact networks

George Cybenko[1] and Steve Huntsman[2*]

*Correspondence:
steve.huntsman@baesystems.com
[2]BAE Systems FAST Labs, 4301 North Fairfax Drive, Arlington, VA 22203, VA, USA
Full list of author information is available at the end of the article

Abstract

Directed contact networks (DCNs) are temporal networks that are useful for analyzing and modeling phenomena in transportation, communications, epidemiology and social networking. Specific sequences of contacts can underlie higher-level behaviors such as *flows* that aggregate contacts based on some notion of semantic and temporal proximity. We describe a simple inhomogeneous Markov model to infer flows and taint bounds associated with such higher-level behaviors, and also discuss how to aggregate contacts within DCNs and/or dynamically cluster their vertices. We provide examples of these constructions in the contexts of information transfers within computer and air transportation networks, thereby indicating how they can be used for data reduction and anomaly detection.

Introduction

Directed contact networks (DCNs) are temporal networks in which edges are directed (Holme 2015; Masuda and Lambiotte 2016). Loosely speaking, temporal networks have time attributes associated with edges while directed contact networks have both time and direction attributes associated with edges. DCNs are a natural temporal generalization of digraphs, and we can think about them informally as collections of time-stamped and directed contacts between and among the entities represented by the nodes. A simple example of a directed contact network is a collection of call data records in which each record includes information about who placed the call, who received the call and the time of the call (Bianchi et al. 2016).

In this paper, we address the problem of how directed contacts can be aggregated and coarsened for purposes such as anomaly detection. To accomplish this, we construct a natural inhomogeneous (that is, time varying) Markov model (Huntsman 2018a) for probabilistic modeling of potential *flows* that aggregate contacts based on a simple notion of spatiotemporal proximity. This model involves a single parameter, which in practice we set automatically with an intuitive heuristic. Through analytical and practical examples, we illustrate the behavior of this Markov model. We emphasize that this Markov model is not statistical in the sense that it involves no learning, fitting, optimization or other estimation procedure. Instead, it starts from a small number of symmetry and invariance requirements that *any* model with its goals ought to obey. This follows a tradition in physics by exhibiting a general mathematical structure that is consistent with the required symmetries. Because of this generality, the model applies to a wide range of problems that can be modeled with directed contacts, including call data record analysis, network traffic analysis and disease surveillance.

We also introduce the concept of a "taint bound" that quantifies the impact of weighted contacts: for example, how much of a given information transfer can possibly propagate through the network. Using publicly available data on flight timetables, we demonstrate by analogy how such taint bounds can constrain data exfiltration within a computer system and network. Finally, we also discuss two ways of aggregating and coarsening networks of directed contacts through renormalization and clustering.

It is useful to note that a contact of any type, information or not, involves a *source* s, a *target* t, and a *time* τ associated with the contact. For simplicity, we assume contacts occur at a given instant in time with the resulting notion of a directed contact as an ordered triple (s, t, τ) . This still enables considerable generality: for instance, a transfer from s to t over the time interval $[\tau_0, \tau_1]$ can be represented by two contacts involving a surrogate third node as $(s, *, \tau_0)$ and $(*, t, \tau_1)$ where $*$ is the surrogate placeholder for $(s, t, [\tau_0, \tau_1])$.[1]

The paper is structured as follows: "Directed contact networks and temporal digraphs" section introduces directed contact networks and temporal digraphs; "Markov chain models for DCNs" section discusses our Markov model, and "Data reduction and anomaly detection" section discusses its performance in data reduction and anomaly detection. We then turn to taint bounds in "Taint bounds" section before discussing renormalization and clustering in directed contact networks in "Renormalization" section and "Clustering" section, respectively. "Remarks" section concludes the paper.

Directed contact networks and temporal digraphs

A particularly useful family of *temporal networks* are *directed contact networks (DCNs)* (Holme 2015; Masuda and Lambiotte 2016). DCNs are a natural temporal generalization of digraphs, and we can think about them informally as collections of contacts as introduced in "Introduction" section. However, to avoid certain degenerate cases, we provide a slightly more formal and restrictive notion here.

A DCN with vertices $V = [n] := \{1, \ldots, n\}$ is a finite nonempty set $\mathcal{C}$, where each *contact* $c \in \mathcal{C}$ corresponds to a unique triple $(s(c), t(c), \tau(c)) \in [n] \times [n] \times \mathbb{R}$ with $s(c) \neq t(c)$. As a matter of convenience, we identify contacts with their corresponding triples in this manner.

Next, define the *temporal fiber at* v, $\mathcal{C}@v$, to be the set of times at which vertex v is involved in a contact as either source or destination, together with the times plus and minus infinity. More formally,

$$\mathcal{C}@v := \{\pm\infty\} \cup \{\tau(c) : c \in \mathcal{C} \wedge (s(c) = v \vee t(c) = v)\} =: \{\tau_j^{@v}\}_{j=0}^{|\mathcal{C}@v|-1}. \tag{1}$$

The *temporal digraph* of $\mathcal{C}$ (for an example, see Fig. 1) is defined as the digraph $T(\mathcal{C})$ with respective vertex and arc sets

$$\begin{aligned} V(T(\mathcal{C})) &:= \{(v, \pm\infty) : v \in V\} \\ &\quad \cup \{(v, \tau(c)) : [(v, c) \in V \times \mathcal{C}] \wedge [s(c) = v \vee t(c) = v]\} \end{aligned} \tag{2}$$

$$\begin{aligned} A(T(\mathcal{C})) &:= \{((s(c), \tau(c)), (t(c), \tau(c))) : c \in \mathcal{C}\} \\ &\quad \cup \{((v, \tau_{j-1}^{@v}), (v, \tau_j^{@v})) : v \in V, j \in [|\mathcal{C}@v| - 1]\}. \end{aligned} \tag{3}$$

[1] By analogy, consider a flight departing from s at τ_0 and arriving at t at τ_1. Here the contact $(s, *, \tau_0)$ corresponds to embarking, while the contact $(*, t, \tau_1)$ corresponds to debarking. We can think of $*$ as the physical plane on which the passengers flew. Alternative representations involving additional contacts of the form $(s, *, \tau_*)$ with $\tau_0 \leq \tau_* < \tau_1$ might also be appropriate depending on circumstances and model intent.

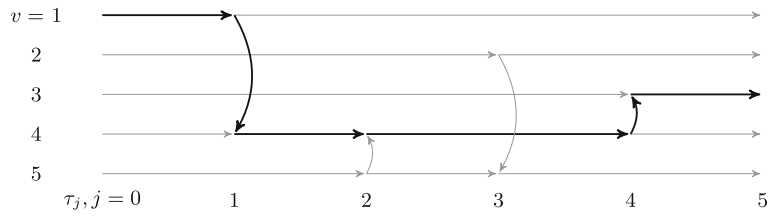

Fig. 1 Temporal digraph of the DCN $\mathcal{C} := \{(1,4,\tau_1),(5,4,\tau_2),(2,5,\tau_3),(4,3,\tau_4)\}$ with $\tau_1 < \tau_2 < \tau_3 < \tau_4$. We indicate a temporally coherent path from $\mathcal{C}$-vertices 1 to 3 with bold versus gray arrows: note that there is no temporally coherent path from 2 to 3. In the figure, spatial (respectively, temporal) arcs are vertical (horizontal). The vertices along a horizontal path constitute the temporal fiber $\mathcal{C}@v$ for v

The first and second sets in the union on the right hand side of (3) are respectively the sets of *temporal arcs* and *spatial arcs*. Because $|V(T(\mathcal{C}))| = \sum_v |\mathcal{C}@v| \leq 2|V| + 2|\mathcal{C}|$ and $|A(T(\mathcal{C}))| = |V(T(\mathcal{C}))| - |V| + |\mathcal{C}| \leq |V| + 3|\mathcal{C}|$, it is easy to see that $T(\mathcal{C})$ can be formed with only linear runtime and memory, though an efficient algorithm requires somewhat more care in practice than is readily apparent.

We also note that DCNs support the natural notion of a *temporally coherent path* based on a set of contacts of the form $\{(s_i, t_i, \tau_i) | s_{i+1} = t_i \ , \ \tau_i \leq \tau_{i+1}\}$. Fig. 1 illustrates a temporally coherent path.

In words, the temporal digraph of a given DCN can be drawn with the horizontal axis representing time and the vertical axis representing nodes in the original DCN in some ordering. Nodes in the temporal digraph are comprised of the start and end point nodes of individual contacts in the DCN at the associated times. As depicted, each vertical arc/edge in the temporal digraph represents a directed contact between two underlying DCN nodes at the specified time while horizontal edges connect a DCN node between the times it is involved in a contact. Note that there are no arcs that go backwards in time in this representation so that all paths are basically from left to right with possible vertical arcs.

Markov chain models for DCNs

In this section, we show that a useful probabilistic model of temporally coherent paths can easily be constructed from $T(\mathcal{C})$ alone. The basic idea comes from traffic analysis, where tools such as pen registers or trap and trace devices generate data that enable a user to make substantive inferences about communication sources and paths in networks (Bianchi et al. 2016).

Specifically, consider two contacts of the form $(A, B, 0)$ and (B, C, τ). A natural probability model for a "flow" from A to C should decrease from 1 to 0 as $\tau \uparrow \infty$. That is, if A calls B and then B calls C, immediately after, there should be a high expectation that some information from A triggered the call to C but if much time transpires between the calls, there is a lower expectation that A's call and communicated information triggered B's call to C.

In practice, we expect that enough flows of interest will involve unusual sources/targets and/or temporally localized contacts to be detected against a background of "bulk traffic" that the model will also effectively characterize. For example, in "Data reduction and anomaly detection" section we show that even sophisticated malicious cyber-activity

leading to so-called "low and slow" data exfiltration involves at least some system call-scale directed contacts that can be readily detected through temporally coherent path identification and analysis.

A reasonable model assigning probabilities to arcs of $T(\mathcal{C})$ should be highly constrained by several fundamental symmetries and invariances that it should obey. We identify four such natural symmetries and invariances:

i) probabilities on spatial arcs from the same source and time should be identical;
ii) the model should yield probabilities for flows that coherently compose over arbitrary consecutive time windows that span the same interval;
iii) probabilities on temporal arcs should only depend on their duration and the number of spatial arcs occurring with the same source and initial time;
iv) simultaneous or near simultaneous events' corresponding probabilities should differ only infinitesimally if at all.

We describe such a physically inspired Markov model of temporally coherent random paths that essentially satisfies these properties. We expect that these properties will be reasonably evident to the mathematically inclined reader.

Define the *restriction* of a DCN $\mathcal{C}$ to $X \subset \mathbb{R}$ to be the subset of contacts with times in X, so that $\mathcal{C}|_X := \tau^{-1}(\tau(\mathcal{C}) \cap X)$. Next, for $a_1 \notin \tau(\mathcal{C})$ and $a_0 < a_1$, we define the *restricted temporal digraph* $T(\mathcal{C})|_{[a_0,a_1)}$ from $T(\mathcal{C}|_{[a_0,a_1)})$ by replacing the time component of the vertices $(v, -\infty)$ with a_0 and the time component of the vertices (v, ∞) with a_1, while retaining all the arcs.

Our probabilistic model also involves an "inverse temperature" $\beta \in \mathbb{R}$ to balance between temporal and spatial arcs.[2] In detail, for $a_0 < \cdots < a_M$ with $a := \{a_m\}_{m=0}^{M}$, $a \cap \tau(\mathcal{C}) = \varnothing$, and $m \in [M]$, we define a Markov chain on the vertices/nodes/states in $V(T(\mathcal{C})|_{[a_{m-1},a_m)})$ to have the transition matrix $P^{(\beta)}_{(\mathcal{C},a,m)}$ given by

$$Z_{(v,\tau_j^{@v})} \cdot P^{(\beta)}_{(\mathcal{C},a,m)}((v,\tau_j^{@v}),(w,\tau_k^{@w})) := \tag{4}$$

$$\begin{cases} 1 & \text{if}[\, v \neq w] \wedge [\, \tau_j^{@v} = \tau_k^{@w}] \\ \exp(-\beta[\, \tau_{j+1}^{@v} - \tau_j^{@v}]\,) & \text{if}[\, v = w] \wedge [j+1 = k] \wedge [\, d^+_{(v,\tau_{j+1}^{@v})} > 0] \\ \exp(-\beta[\, \tau_{(a,m)}^{@v+} - \tau_j^{@v}]\,)) & \text{if}[\, v = w] \wedge [j+1 = k] \wedge [\, d^+_{(v,\tau_{j+1}^{@v})} = 0] \\ 1 & \text{if}[\, v = w] \wedge [j = k] \wedge [\, d^+_{(v,\tau_j^{@v})} = 0] \\ 0 & \text{otherwise.} \end{cases}$$

where $\tau_{(a,m)}^{@v+} := \min(\inf[\,(a_m,\infty) \cap \mathcal{C}@v]\,, a_M)$, the normalizing constants $Z_{(v,\tau_j^{@v})}$ are such that the rows of $P^{(\beta)}_{(\mathcal{C},a,m)}$ sum to 1, and d^+ denotes the outdegree in $T(\mathcal{C})|_{[a_{m-1},a_m)}$. The reader should note that terms of the form $\exp(-\beta\Delta\tau)$ can easily underflow numerically if the exponential argument is large and negative so that care must be taken when implementing these formulae in finite precision arithmetic.

Figure 2 shows a simple example of this definition overlaid on the example of Fig. 1.

[2] Allowing a negative absolute temperature (Ramsey 1956), $\beta = -\infty$ and $\beta = \infty$ respectively correspond to "absolute hot" (no spatial arc traversals) and absolute zero (no temporal arc traversals). **In practice, we use a physical analogy/heuristic to set β^{-1} to the average time between contacts. Further discussion of the role of β can be found below.**

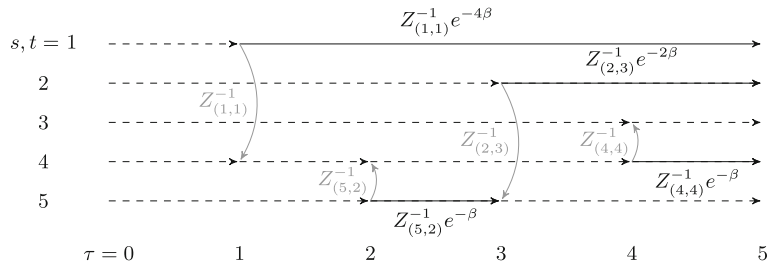

Fig. 2 Entries of $P^{(\beta)}_{(\boldsymbol{\mathcal{C}},a',1)}$ not in $\{0,1\}$ for $\boldsymbol{\mathcal{C}} := \{(1,4,1),(5,4,2),(2,5,3),(4,3,4)\}$ and $a' = \{0,5\}$ are indicated along with solid arcs (gray for spatial arcs; black for temporal arcs); unit entries correspond to dashed arcs

The specific formulae of (4) arise from the four identified constraints above rather than from arbitrary choices.

In particular, the requirement $a \cap \tau(\mathcal{C}) = \varnothing$ ensures that the Markov chain defined by (4) has exactly n absorbing states. Each of these has the form (v, a_m) and a corresponding "emitting" state (v, a_{m-1}), and a natural quantity to consider is the probability $\mathcal{Q}^{(\beta)}_{(\mathcal{C},a,m)}(v,w)$ of arriving at the absorbing state (w, a_m) after starting in the emitting state (v, a_{m-1}). We can straightforwardly compute this quantity using the so-called *fundamental matrix* (Brémaud 1999). Finally, the term $\tau^{@v+}_{(a,m)}$ mitigates artificial "boundary effect" behavior for $m = M$.

Taken as a whole, (4) thus leads to a natural temporal coherence property and a straightforward physical interpretation. Regarding the symmetries and invariants listed above, the terms equal to 1 and 0 in (4) merely codify i), while ii) is embodied in the following identity which can easily be verified from the construction of the probability transition matrices, $\mathcal{Q}^{(\beta)}_{(\mathcal{C},a,m)}$.

Lemma 1 *If* $a_0, a_{|a|-1} \in a' \subseteq a$, *then*

$$\mathcal{Q}^{(\beta)}_{(\mathcal{C},a',1)} \cdot \cdots \cdot \mathcal{Q}^{(\beta)}_{(\mathcal{C},a',|a'|-1)} = \mathcal{Q}^{(\beta)}_{(\mathcal{C},a,1)} \cdot \cdots \cdot \mathcal{Q}^{(\beta)}_{(\mathcal{C},a,|a|-1)}. \tag{5}$$

Proof The lemma follows by applying the Kolmogorov-Chapman property to each of the inhomogeneous Markov chains $\mathcal{Q}^{(\beta)}_{(\mathcal{C},a',\cdot)}$ and $\mathcal{Q}^{(\beta)}_{(\mathcal{C},a,\cdot)}$. □ □

That is, for any DCN $\mathcal{C}$ and parameter β, we have an associated temporally coherent family of time-inhomogeneous Markov chains. This lemma does not rely on the specific form of (4): the $\exp(-\beta \cdot \Delta\tau)$ terms could be significantly changed without breaking property ii) above.[3]

Moreover, this form is necessary to jointly satisfy properties iii) and iv): i.e., memorylessness and self-consistency in the limit $\Delta\tau \downarrow 0$. (In particular, the self-consistency requirement prohibits multiplying the exponentials by some nontrivial constant.) That is, the form of (4) is dictated by the structure of a temporal digraph along with manifestly desirable symmetries.[4]

[3] That said, a dependence on $\Delta\tau$ is necessary. In the context of intra-computer information flows, this time difference plausibly approximates (at least for small values) a linear function of the conditional Kolmogorov complexity of the intervening computation.

[4] We can generalize this construction to the related notion of a weighted DCN by normalizing the sum of outbound weights and modifying the first case in (4) accordingly.

In the limit $\beta \to \infty$, the trajectories of the Markov model are so-called *greedy walks* (Saramäki and Holme 2015), and more generally for $\beta > 0$, "spatial" transitions are preferred over "temporal" transitions. Regardless of the value of β, we shall see that the temporal coherence of trajectories is captured more faithfully and conservatively in the Markov model than in series of "projected snapshots" that characterize earlier efforts such as Perra et al. (2012); Starnini et al. (2012); Rocha and Masuda (2014); Grindrod and Higham (2013); Valdano et al. (2015) to analyze DCNs through graph time series and/or provide a substrate for random walks. A recent notable work that develops techniques for special epidemiological models can be found in Valdano et al. (2018).

The preceding lemma facilitates a computational complexity analysis as a function of a. Writing $N := |\mathcal{C}|$ and $M := |a|$, we suppose that a is approximately uniform in the sense that $\left|\mathcal{C}|_{[a_m,a_{m+1})}\right| \approx N/M$, which also implies that $|V(T(\mathcal{C}|_{[a_m,a_{m+1})}))| \approx 2(n + N/M)$. The complexity of computing $\mathcal{Q}^{(\beta)}_{(\mathcal{C},a,m)}$ is governed by a matrix division of the form $(I - Q)\backslash R$, where here Q is the block of $P^{(\beta)}_{(\mathcal{C},a,m)}$ whose rows and columns both correspond to transient states, and R is the block whose rows and columns respectively correspond to transient and absorbing states. Since the numbers of transient and absorbing states are respectively approximately $n + 2N/M$ and exactly n, the complexity of computing $(I - Q)\backslash R$ is $O(n(n + 2N/M)^{\omega-1})$, where we take matrix multiplication and inversion to have complexity exponent $\omega > 2$ (for dense unstructured matrices, in practice $\omega = 3$). Because there are $M - 1$ matrix multiplications, the computational complexity for the right hand side of (5) is $O(Mn(n + 2N/M)^{\omega-1})$. Now $\arg\min_M Mn(n + 2N/M)^{\omega-1} = 2(\omega - 2)N/n$, and this value for M yields computational complexity which is nominally linear in N. Meanwhile, it only makes sense to take $M \ll N/n$ if the complexity of the linear algebra involved is dominated by the rest of the computation. In other words, it is less expensive to invert and multiply many small matrices than to invert and multiply a few large matrices. Since taking M larger yields a more detailed picture of the dynamics of $\mathcal{C}$, it is sensible to require M to be (at least) on the order of N/n.

We exhibit the basic mechanics of the model in the following

Example 1 *Consider once more the DCN shown in Fig. 1. Let $\tau_j = j$ for $1 \le j \le 4$, $\varepsilon \ll 1$, $a = \{0, 2.5, 5\}$ and $a' = \{0, 5\}$. The entries of $P^{(\beta)}_{(\mathcal{C},a',1)}$ are then as in Fig. 2 and (using $\cdot$ in matrices to denote 0 for clarity)*

$$
\begin{aligned}
\mathcal{Q}^{(\beta)}_{(\mathcal{C},a',1)} &= \begin{pmatrix}
\frac{e^{\beta}+1}{(e^{4\beta}+1)(e^{\beta}+1)} & \cdot & \frac{e^{5\beta}}{(e^{4\beta}+1)(e^{\beta}+1)} & \frac{e^{4\beta}}{(e^{4\beta}+1)(e^{\beta}+1)} & \cdot \\
\cdot & \frac{1}{e^{2\beta}+1} & \cdot & \cdot & \frac{e^{2\beta}}{e^{2\beta}+1} \\
\cdot & \cdot & 1 & \cdot & \cdot \\
\cdot & \cdot & \frac{e^{\beta}}{e^{\beta}+1} & \frac{1}{e^{\beta}+1} & \cdot \\
\cdot & \cdot & \frac{e^{2\beta}}{(e^{\beta}+1)^2} & \frac{e^{\beta}}{(e^{\beta}+1)^2} & \frac{e^{\beta}+1}{(e^{\beta}+1)^2}
\end{pmatrix} \\
&= \begin{pmatrix}
\frac{1}{e^{4\beta}+1} & \cdot & \cdot & \frac{e^{4\beta}}{e^{4\beta}+1} & \cdot \\
\cdot & 1 & \cdot & \cdot & \cdot \\
\cdot & \cdot & 1 & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & \cdot \\
\cdot & \cdot & \cdot & \frac{e^{\beta}}{e^{\beta}+1} & \frac{1}{e^{\beta}+1}
\end{pmatrix} \cdot \begin{pmatrix}
1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \frac{1}{e^{2\beta}+1} & \cdot & \cdot & \frac{e^{2\beta}}{e^{2\beta}+1} \\
\cdot & \cdot & 1 & \cdot & \cdot \\
\cdot & \cdot & \frac{e^{\beta}}{e^{\beta}+1} & \frac{1}{e^{\beta}+1} & \cdot \\
\cdot & \cdot & \cdot & \cdot & 1
\end{pmatrix} \\
&= \mathcal{Q}^{(\beta)}_{(\mathcal{C},a,1)} \cdot \mathcal{Q}^{(\beta)}_{(\mathcal{C},a,2)},
\end{aligned}
\tag{6}
$$

so that with increasing β (or equivalently, decreasing temperature) the likeliest transitions correspond to temporally coherent paths that greedily traverse spatial arcs. Now consider the digraph D with arcs $(s(c), t(c))$ for $c \in \mathcal{C}$ and loops (v, v) for $v \in [n]$. The adjacency matrix of D has nonzero entries in the same locations as $\mathcal{Q}^{(\beta)}_{(\mathcal{C},a',1)}$, and also in the $(2,3)$ and $(2,4)$ locations. The $(2,3)$ and $(2,4)$ entries correspond to spurious temporally coherent paths in $\mathcal{C}$. In particular, $\mathcal{Q}$ gives a more faithful description of $\mathcal{C}$ than D. □

Although the context is quite different, the closest work to the construction of this section is Ser-Giacomi et al. (2015), which shows that the most probable paths in a Markovian model of a very complicated temporal network (viz., ocean water transport in the Mediterranean) suffice to describe the network's key features. Other works have looked at higher-order models in discrete time as a way to finesse the challenges of continuous time modeling as discussed here (Lambiotte et al. 2019; Rosvall et al. 2014). Despite the many differences of detail, our own model likewise shows that the most probable paths/flows suffice for capturing the essential dynamics of directed contact networks. In particular, this includes flows that the model assesses as highly probable, but whose associated contact motifs occur infrequently (or perhaps just once in a given data set): in our experiments, such flows reliably capture anomalous and even malicious behavior (see, for example "Data reduction and anomaly detection" section).

Embeddability

It is natural to wonder under what if any circumstances the Markov chain $\mathcal{Q}^{(\beta)}_{(\mathcal{C},a,m)}$ corresponds to a continuous-time Markov process. As it happens, this instance of the *Markov embeddability* problem Lencastre et al. 2016) can be answered quite effectively (if not always affirmatively) for most situations of practical interest.

In the event that no two contacts are simultaneous, this problem reduces to the case of a single contact for $n = 2$, which in turn follows from the following identity for $p \in (0, 1)$, which can be verified by using the power series expansion for log:

$$\log \begin{pmatrix} 1-p & p \\ 0 & 1 \end{pmatrix} = \log(1-p) \cdot \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}. \tag{7}$$

On the other hand, simultaneous contacts are a possible obstruction to embeddability. For $n = 2$, a stochastic matrix is embeddable iff it has positive determinant. But a quick calculation for $\mathcal{C} := \{(1,2,0), (2,1,0), (1,2,\tau), (2,1,\tau)\}$ and $a = \{-1, \tau/2, 2\tau\}$ shows that $\det \mathcal{Q}^{(\beta)}_{(\mathcal{C},a,1)} < 0$ for $\beta > 0$.

Proposition IV.3 of Lencastre et al. (2016) immediately yields a generalization of the preceding observations:

Proposition 1 *If $T(\mathcal{C})$ is acyclic, then $\mathcal{Q}^{(\beta)}_{(\mathcal{C},a,m)}$ is embeddable.* □

Finally, when a Markov generator exists, the algorithm of §V.B of Lencastre et al. (2016) can be used to estimate it.

Data reduction and anomaly detection

In this section, we sketch how the model of "Markov chain models for DCNs" section performs data reduction well enough to be used as a practical anomaly detector. For some additional background details of this analysis, see (Huntsman 2018b).

We considered a DCN $\mathcal{C}$ formed from $N \approx 3.4 \cdot 10^6$ kernel-level events spanning a period of four days and derived in turn from data produced by the CADETS tool (Jenkinson et al. 2017). Also, after curation, we obtained a set $\mathcal{G} \subset \mathcal{C}$ of 216 ground truth contacts that were distributed over 54 malicious exfiltration events.

At a high level, we mapped an event represented as

(*timestamp*, (*process name*, *process identifier*), *event type*, *filename*)

(where the "everything is a file" philosophy applies to the last entry, such as, for a fork event, the filename is the forked process identifier) onto

(*process name*, *filename*, *timestamp*),

or

(*filename*, *process name*, *timestamp*),

or both, depending on the semantics of *event type*. While many *event type* semantics included a natural and unambiguous (bi)directionality determining the mapping above, some did not, and in that case both contacts were conservatively included.

We obtained flows using the model of "Markov chain models for DCNs" section by setting β to the mean inter-contact time (i.e., the proposed heuristic) and used $M = 28423$ windows of 10 s.

We let $\hat{\mathcal{I}}(m)$ denote the set of indices corresponding to sources or targets of flows spanning the mth time window that simultaneously fell above a flow probability threshold $\lambda \in [0, 1]$ and below a per-window frequency threshold $\mu \in [0, 1]$. The set $\hat{\mathcal{I}}(m)$ is an estimate of the set $\mathcal{I}(m)$ of indices corresponding to sources or targets of ground truth events during the mth time window.[5]

Using the estimated and ground truth index sets $\hat{\mathcal{I}}(m)$ and $\mathcal{I}(m)$, we defined two versions of detection metrics. Suppressing the argument m for clarity, the "Boolean" version uses $\mathit{True} = \top := 1$ and $\mathit{False} = \bot := 0$ so that

$$\begin{aligned}
\delta_{\text{bool}}^{\top+} &:= \left[\hat{\mathcal{I}} \neq \varnothing\right] \wedge \left[\hat{\mathcal{I}} \cap \mathcal{I} \neq \varnothing\right]; \\
\delta_{\text{bool}}^{\bot+} &:= \left[\hat{\mathcal{I}} \neq \varnothing\right] \wedge \left[\hat{\mathcal{I}} \cap \mathcal{I} = \varnothing\right]; \\
\delta_{\text{bool}}^{\bot-} &:= \left[\hat{\mathcal{I}} = \varnothing\right] \wedge \left[\mathcal{I} \neq \varnothing\right]; \\
\delta_{\text{bool}}^{\top-} &:= \left[\hat{\mathcal{I}} = \varnothing\right] \wedge \left[\mathcal{I} = \varnothing\right]. \qquad (8)
\end{aligned}$$

The natural number analogues of (8) are

$$\delta_{\text{nat}}^{\top+} := \left|\hat{\mathcal{I}} \cap \mathcal{I}\right|; \quad \delta_{\text{nat}}^{\bot+} := \left|\hat{\mathcal{I}} \cap \mathcal{I}^c\right|; \quad \delta_{\text{nat}}^{\bot-} := \left|\hat{\mathcal{I}}^c \cap \mathcal{I}\right|; \quad \delta_{\text{nat}}^{\top-} := \left|\hat{\mathcal{I}}^c \cap \mathcal{I}^c\right|. \qquad (9)$$

From these we get in turn the usual detection metrics shown in Figs. 3 and 4, i.e. true positive rate (or recall) and false positive rate

[5] This construction was necessary because in many cases the source or target of a ground truth event did not exist. For example, the userspace commands `hostname` and `put /tmp/netrecon` correspond to the (*process name*, *filename*) pairs (`hostname`, $\varnothing$); and ($\varnothing$, `/tmp/netrecon`). By way of comparison, the command `rm -f /tmp/netrecon.log` corresponds to the pair (`rm`, `/tmp/netrecon.log`).

$$\text{TPR} := \frac{\sum_m \delta^{\top+}(m)}{\sum_m \delta^{\top+}(m) + \sum_m \delta^{\perp-}(m)}; \quad \text{FPR} := \frac{\sum_m \delta^{\perp+}(m)}{\sum_m \delta^{\perp+}(m) + \sum_m \delta^{\top-}(m)}, \tag{10}$$

and positive predictive value (or precision) and negative predictive value

$$\text{PPV} := \frac{\sum_m \delta^{\top+}(m)}{\sum_m \delta^{\top+}(m) + \sum_m \delta^{\perp+}(m)}; \quad \text{NPV} := \frac{\sum_m \delta^{\top-}(m)}{\sum_m \delta^{\perp-}(m) + \sum_m \delta^{\top-}(m)}. \tag{11}$$

From Figs. 3 and 4, we can see that the results were insensitive to the probability threshold λ.[6] Similarly, a cursory analysis indicated broad insensitivity to the value of β over several orders of magnitude, a fact attributable to information flow probabilities that tended to be either very near or bounded away from 1. This also underlies the insensitivity with respect to the probability threshold λ. For the value $\mu = 10^{-3}$, a majority of malicious events were detected with a false positive rate below 2 percent (by either version of the metrics).

The results indicate that the Markov model is a sufficiently effective data reduction technique (in particular, the negative predictive value is essentially perfect) to be a useful anomaly detector. In fact, of the 57 (out of 418) files which are targets of high-probability potential information flows in the model, 27 fell below the $\mu = 10^{-3}$ level and had backtracks (King and Chen 2005) with fewer than 20 (or for that matter, 90) vertices. From these 27, 6 (in 3 pairs) corresponded to the 3 executables which the malicious attacker wrote to `/tmp` from its initial foothold.

Taint bounds

The notions of *dynamic taint analysis* (Schwartz et al. 2010) and *provenance* (Cheney et al. 2013) inform the context where a DCN models information flow in a computational environment. The analytic problems corresponding to these notions are generically undecidable. With this in mind, we introduce the idea of *taint bounds*, wherein correct nontrivial bounds on the information flow are maintained.[7] We formalize this idea here before showing its utility as a practical guide to producing effective data-reducing path abstractions.

Let ρ be a nontrivial binary relation on a finite set X such that the transitive closure ρ^+ is irreflexive (such relations have been called *superirreflexive* (Flaška et al. 2007)), and hence also a strict partial order. Let $\gamma : X \to [0, \infty)$ and define the *lower taint bound* α and *upper taint bound* β for $x \in X$ as follows:

$$\alpha(x) := \left(\bigwedge_{x_1 \rho x} \alpha(x_1) \right) \wedge \gamma(x); \tag{12}$$

$$\beta(x) := \left(\sum_{x_1 \rho x} \beta(x_1) \right) \wedge \gamma(x) \tag{13}$$

where $a \wedge b = \min(a, b)$ in this section. Here we note that the standard interpretation here is that not only minima but also summation over the empty set yield ∞.

[6] In more delicate situations, the approach of Huntsman (2018a) offers a principled solution to the problem of thresholding.

[7] Notwithstanding their fundamentally dynamic character, these bounds may be regarded as having a loose analogue in the practice of *abstract interpretation* in static analysis of computer programs (Nielson et al. 2010).

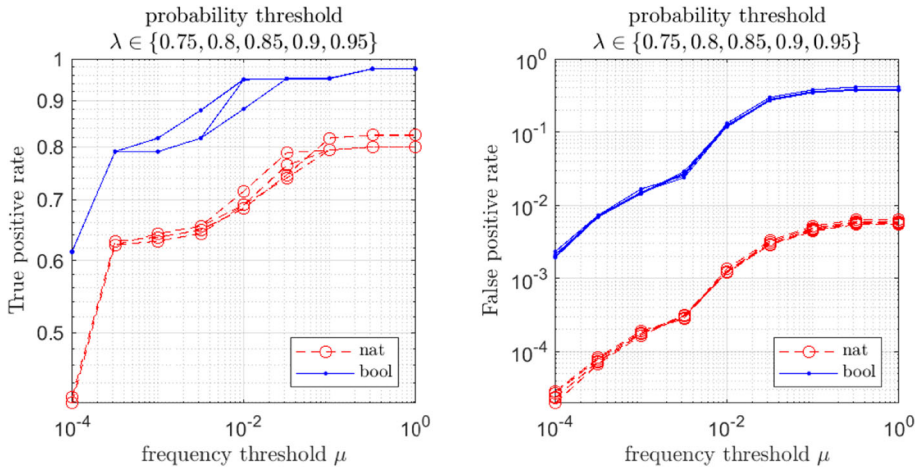

Fig. 3 True (left) and false (right) positive rates as defined in (10) for APT detections using the model of "Markov chain models for DCNs" section and as functions of a per-window frequency threshold μ and for varying flow probability thresholds λ

Lemma 2 *α and β are well-defined; moreover, $\alpha \leq \beta \leq \gamma$ and*

$$\alpha(x) = \left(\bigwedge_{x' \rho^+ x} \gamma(x') \right) \wedge \gamma(x). \tag{14}$$

Proof By lemma 3.5 of Flaška et al. (2007), a binary relation ρ is superirreflexive iff the digraph corresponding to ρ is acyclic. Since by assumption ρ is nontrivial and X is finite, there exists some $j < \infty$ such that $\rho^{\circ(j+1)} = \varnothing$ and $\rho^{\circ j} \neq \varnothing$, where the composition of ρ with itself is indicated. Furthermore, superirreflexivity implies irreflexivity.

Therefore, the recursion implicit in (12) terminates and we have

$$\alpha(x) = \left[\bigwedge_{x_1 \rho x} \left(\left[\bigwedge_{x_2 \rho x_1} \cdots \left(\left[\bigwedge_{x_{j+1} \rho x_j} \alpha(x_{j+1}) \right] \wedge \gamma(x_j) \right) \cdots \right] \wedge \gamma(x_1) \right) \right] \wedge \gamma(x). \tag{15}$$

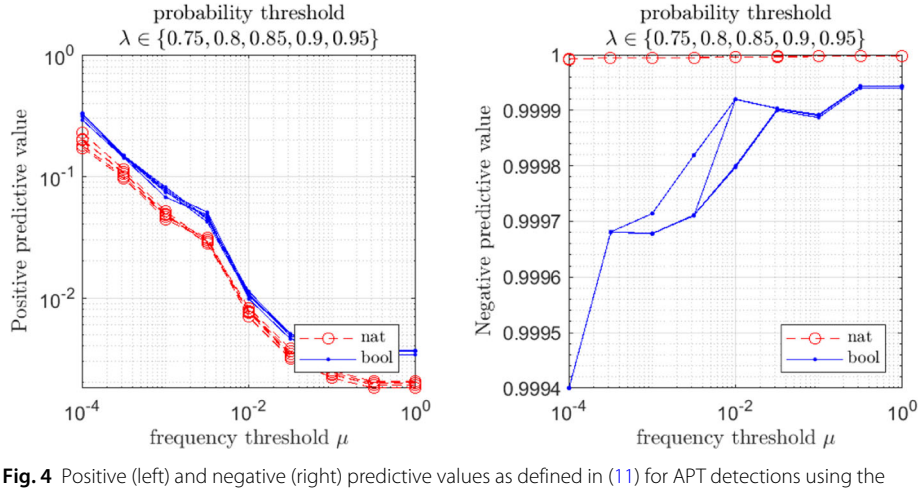

Fig. 4 Positive (left) and negative (right) predictive values as defined in (11) for APT detections using the model of "Markov chain models for DCNs" section and as functions of a per-window frequency threshold μ and for varying flow probability thresholds λ

But $\bigwedge_{x_{j+1}\rho x_j} \alpha(x_{j+1}) = \infty$ and $x_k \rho^{\circ k} x$ for all $k \in [j]$, so

$$\alpha(x) = \left(\bigwedge_{x_1 \rho^{\circ j} x} \gamma(x_j) \right) \wedge \cdots \wedge \left(\bigwedge_{x_1 \rho^{\circ 1} x} \gamma(x_1) \right) \wedge \gamma(x) = \left(\bigwedge_{x' \rho^{+} x} \gamma(x') \right) \wedge \gamma(x). \quad (16)$$

Similarly, the recursion implicit in (13) also terminates, though in this case without any additional simplification. The bounds $\alpha \leq \beta \leq \gamma$ follow. □ □

For a DCN $\mathcal{D}$ such that τ is injective, a natural choice for ρ is $c_1 \rho c \iff (t(c_1) = s(c)) \wedge (\tau(c_1) < \tau(c))$. Note that this relation is superirreflexive.

Example 2 *Consider once more the DCN depicted in Fig. 1 and ρ as defined immediately above. The only nontrivial taint bounds are for the last contact: we have that $\alpha(4, 3, \tau_4) = (\gamma(1, 4, \tau_1) \wedge \gamma(5, 4, \tau_2)) \wedge \gamma(4, 3, \tau_4)$ and $\beta(4, 3, \tau_4) = (\gamma(1, 4, \tau_1) + \gamma(5, 4, \tau_2)) \wedge \gamma(4, 3, \tau_4)$. If for $3 \leq j \leq J$ we add to this DCN the contacts $(3, 4, \tau_{2j-1})$ and $(4, 3, \tau_{2j})$ with τ_k strictly increasing, then the result is a DCN with $2J$ contacts and $\alpha(\cdot, \cdot, \tau_k)$ and $\beta(\cdot, \cdot, \tau_k)$ nonincreasing for $k > 4$.* □

However, in practice τ need not be injective, and in fact this is often the case for kernel-level information flows (timestamps of system-level activity in computers or network interfaces are generally precise only to milliseconds or at best microseconds, and getting higher precision generally entails a heavy burden or can even be practically infeasible, depending on the detailed context). Indeed, for a distributed system, even the synchronization of clocks can become an issue, and so it is desirable to have a relation ρ that accounts for more structural details. The problem is highlighted by considering directed acyclic graphs (DAGs) rather than DCNs: for a DAG D, the natural choice for ρ is $u \rho v$ iff *u precedes v*. However, this does not generalize to arbitrary digraphs, which are the structures that essentially embody multiple contacts occurring at the same time.

This suggests two strategies: live with a fairly generic relation ρ and only seek to compute taint bounds when the digraph corresponding to ρ actually turns out to be acyclic, or build in mechanisms that enforce acyclicity (if these are artificial, we can provide warnings when they have any effect). Only the second of these strategies requires further comment here. One simple approach is to leverage some auxiliary strict order on X; another simple approach is to require a nonzero delay between contacts. In general, context will constrain and inform the construction of ρ.

Example 3 *Consider the set of $N = 1155$ scheduled commercial nonstop domestic flights in the United Kingdom (UK) and Crown dependencies on Monday, 18 October 2010 (Gallotti and Barthelemy 2015a; Gallotti and Barthelemy 2015b). We form a DCN from these by associating each flight with two contacts: one from the flight's origin to the flight itself, and one from the flight to the flight's destination. We take an aggressive approach to determining connecting flights, viz. $c_1 \rho c$ only if either c_1 and c are respectively the first and second contact corresponding to a single flight, or $(t(c_1) = s(c)) \wedge (\tau(c_1) \leq \tau(c) - 1/2)$, where time is measured in hours (note that the required half-hour layover ensures that ρ is superirreflexive). We take γ to be the number of seats available on a flight. Only the 7 flights in Table 1 (where origin and destination are indicated using International Air Transport Association airport codes) correspond to contacts with $\beta < \gamma$.*

Table 1 Taint bounds with $\beta < \gamma$ for UK flight data in Example 3

Origin	Destination	Seats	Departure	Arrival	β	γ
CWL	EDI	74	0850	1010	29	74
CWL	GLA	74	0850	1010	29	74
BRS	BHD	189	0745	0855	82	189
BLK	BFS	142	1405	1455	40	142
GLO	IOM	18	1025	1130	8	18
INV	LTN	149	1115	1245	142	149
GLA	EMA	136	0835	0945	74	136

Let us consider each of these in turn. On the day in question, one flight arrives at CWL by 0820, and it has 29 seats. Two flights arrive at BRS by 0715, each with 41 seats. Two flights arrive at BLK by 1335, each with 20 seats. One flight arrives at GLO by 0955, and it has 8 seats. Three flights arrive at INV by 1045, with 34, 34, and 74 seats, respectively. Finally, one flight arrives at GLA by 0805, and it has 74 seats. The general pattern amongst these is evident by inspection: there are a few preceding flights inbound to the origin with less total capacity than the current flight. □

The preceding example illustrates by way of analogy how differences between β and γ can serve as a preliminary indicator of anomalously asymmetric flows (which might correspond to, for example, the original dissemination of material and/or unauthorized data exfiltration), particularly at vertices corresponding to sensitive objects or locations.

Renormalization

For $\tau_1 < \cdots < \tau_N \in \mathbb{R}$, consider the DCN $\mathcal{C} := \{(1, 2, \tau_j)\}_{j=1}^{N-1}$. It is of interest to try to associate $\mathcal{C}$ with a single coarse-grained or "renormalized" contact using the Markov model of "Markov chain models for DCNs" section. Writing $\mathcal{Q} = \mathcal{Q}^{(\beta)}_{(\mathcal{C},\{-\infty,\tau_N\},1)}$, we have that

$$\mathcal{Q}_{11} = \prod_j \frac{e^{-\delta_j}}{1 + e^{-\delta_j}}, \tag{17}$$

where $\delta_j := \beta(\tau_{j+1} - \tau_j)$. If we pick Δ so that

$$\mathcal{Q}_{11} =: \frac{e^{-\Delta}}{1 + e^{-\Delta}} \tag{18}$$

then

$$\Delta = \log\left(\prod_j [\, e^{\delta}_j + 1] - 1\right). \tag{19}$$

We want $0 \leq \Delta \leq \sum_j \delta_j$, so that the notional renormalized contact $(1, 2, \tau_N - \Delta)$ can replace $\mathcal{C}$ in a self-consistent way. While the first inequality holds, the second is equivalent to $\prod_j [\, e^{\delta_j} + 1] \leq \prod_j e^{\delta_j} + 1$, which is impossible for $\beta > 0$. That is, our goal of associating $\mathcal{C}$ with a single renormalized contact is generally impossible.

In light of the preceding considerations, it seems necessary to resort to more algorithmical and computational versus analytical approaches to coarse-graining or renormalizing DCNs. At the same time, it is helpful to introduce some additional context. Stripped bare of its associations with physics, the *renormalization group* (RG; see, for example, (Barenblatt 2003; Goldenfeld 1992)) is a simple approach to understanding theories in terms of

their fixed points. We also note that renormalization ideas have been applied to undirected networks with specific structures (Barrat and Cattuto 2013; Karschau et al. 2018; Newman and Watts 1999).

For a theory determined by a function $f(x;\theta)$ of data x and parameters θ, and given a suitable coarsening operator C, if there exists a function g such that $f(x;\theta) = f(C(x);g(\theta))$, then the theory is called *renormalizable*. [8] In our setting, a probability cutoff takes the role of f; the underlying DCN takes the role of x; the parameter $\beta = \theta$ is computed from the data x according to a fixed heuristic; and the coarsening operator C is realized by the Markov model of "Markov chain models for DCNs" section along with a fixed heuristic for its remaining parameters - for instance, we can fix the number of contact times per window (with an exception provided for the last window). The use of fixed heuristics yields a RG transformation on DCNs that renormalizes probable flows into contacts in a given time window.

Iterating the RG transformation along these lines leads to an "ultraviolet cutoff" at which the process stops, essentially sublimating temporal data into a single weighted digraph. While there is a great deal of freedom in its precise specification, such an RG transformation and fixed point is surely of interest for summarizing complex DCNs.

In Table 2, we show the sizes of DCNs obtained through such RG transformations up to a fixed point in an experiment on data similar to that described in "Data reduction and anomaly detection" section. Of the final 184 renormalized contacts, at least 8% appeared to be associated with malicious activity.

Clustering

The problem of clustering in digraphs is much more delicate than its analogue for the undirected case (Malliaros and Vazirgiannis 2013). It should therefore come as no surprise that the problem of clustering in DCNs is more challenging than either clustering in digraphs or in undirected temporal networks. Indeed, most of the approaches purporting to address clustering in temporal networks in the literature (cf. §4.11 of (Holme 2015), §4.12 of Masuda and Lambiotte (2016) or Speidel et al. (2015)) actually cluster in time series of graphs, not the more granular notion of a DCN.

A sensible step forward is to consider the temporal digraph $T(\mathcal{D})$ of a DCN $\mathcal{D}$. As an "almost acyclic" digraph, it might seem natural to try to apply techniques such as those detailed in Malliaros and Vazirgiannis (2013) directly to $T(\mathcal{D})$. While this would offer the prospect of retaining qualitative temporal structure, it still ignores the quantitative temporal details; furthermore, it is far from evident how to remove any cycles that might (and in practice frequently do) occur. We seek instead a controlled way to coarse-grain this temporal information independent of the approach in "Renormalization" section.

We note that clustering for temporal networks is a topic of much current interest (Bassett et al. 2013; Bazzi et al. 2016; Gauvin et al. 2014; Sarzynska et al. 2015) seeing as the dynamics of communities within social networks and other applications are relevant to current social media and related topics. However, our focus here is an on the specific structure of directed contact networks which has not been specifically studied to our knowledge before.

[8] Renormalizable theories in physics (and their fixed/critical points) are of great interest: indeed, renormalizability is actually a requirement for statistical and quantum field theories to be well-defined rather than "effective."

Table 2 Reduction of data similar to that described in "Data reduction and anomaly detection" section under RG transformations

RG iteration	Number of contacts
0	569480
1	10726
2	4687
3 (∞)	184

Clustering techniques leveraging (5)

The time-inhomogeneous Markov chain (5) provides a platform for any number of capabilities, not least clustering. Before plunging ahead, the taxonomy of Malliaros and Vazirgiannis (2013) for digraph clustering suggests some guiding principles:

1. Any clustering technique ought to directly exploit the probabilistic framework that (5) offers and seek to avoid any additional model features unless they are necessary. This principle discourages–but of course does not completely rule out–techniques that require for example random walks generated by an ergodic transition matrix, which would in turn require incorporating a "teleportation" device *à la* PageRank. Techniques that require a unique and/or nondegenerate stationary distribution are therefore also discouraged by this principle. Such discouraged techniques include (following the precise enumeration in table 2 of Malliaros and Vazirgiannis (2013)) symmetrization and random walk simulations, LinkRank, directed Laplacians, two-step random walks, message passing, and Infomap. Meanwhile, many other techniques do not exploit (5) at all and should be completely ruled out: for example, network embedding, bipartite modularity, a modified adaptive genetic algorithm, semi-supervised learning, directed modularity, directed Gaussian random network, overlapping modularity, local modularity, cuts, attraction/repulsion, local partitioning, directed clique percolation, local density, mixture models, and community kernels.
2. The techniques in Malliaros and Vazirgiannis (2013) not discouraged or completely ruled out by the immediately preceding considerations essentially amount to *coclustering* (or the closely related notion of "blockmodels"). Among coclustering approaches, we single out (Ge et al. 2003; Chakrabarti 2004; Rohe et al. 2016) as holding particular interest. (Ge et al. 2003) focuses on reducing the number of states of a Markov chain estimated directly from a sample trajectory (and is thus not manifestly suitable in our context, where the data is a sequence of contacts rather than a sequence of vertices), while (Chakrabarti 2004; Rohe et al. 2016) explicitly address unweighted digraphs. (Ge et al. 2003; Rohe et al. 2016) cluster singular vectors of a suitable matrix, whereas (Chakrabarti 2004) optimizes a minimum description length criterion for coclustering. Bearing all this in mind, a reasonable strategy would be to look for opportunities to evaluate the singular value decomposition or an information-theoretical compression of a suitable matrix. At the same time, the notion of stochastic equivalence (Holland et al. 1983) leveraged by (Rohe et al. 2016) appears particularly relevant: vertices v and w are stochastically equivalent for (5) iff $\mathcal{Q}_{v\cdot} = \mathcal{Q}_{w\cdot}$ and $\mathcal{Q}_{\cdot v}, \mathcal{Q}_{\cdot w}$, where we use a shorthand. We shall exploit a very similar notion immediately below.

Intuitively, a time-dependent clustering for vertices of a DCN that models the flow of some quantity ought to be determined by a metric involving the probabilities of transitions to and from states. Let d denote an arbitrary metric on probability distributions and write $\mathcal{Q}$ for a matrix such as in (5). If we are only concerned with the probabilities of transitions from states, then it suffices to consider

$$d(v, w) := d(\mathcal{Q}_{v\cdot}, \mathcal{Q}_{w\cdot}). \tag{20}$$

If instead we are only concerned with the probabilities of transitions to states, the situation is more delicate. The reason is that in general (5) will not yield a very well-behaved Markov chain, even apart from any time-inhomogeneity. For example, reducibility is common in practice. This means that the classical notion of time reversal for the chain is not well-defined, which complicates any attempt to consider the probabilities of transitions to states. Nevertheless, it is easy to construct an essentially unique time reversal $\mathcal{D}^{\leftarrow}$ of the underlying DCN $\mathcal{D}$ by merely swapping sources and targets and replacing τ with $\tau_* - \tau$ for any fixed $\tau_* \in \mathbb{R}$. Writing $\mathcal{Q}^{\leftarrow}$ for a matrix obtained by applying (5) to $\mathcal{D}^{\leftarrow}$, the time-reversed analogue of (20) is

$$d^{\leftarrow}(v, w) := d(\mathcal{Q}^{\leftarrow}_{v\cdot}, \mathcal{Q}^{\leftarrow}_{w\cdot}). \tag{21}$$

Meanwhile, if we are concerned with the probabilities of transitions both to and from states, it is both natural and easy to consider for $0 < q < \infty$ (with an extension to $q = \infty$) an induced metric (and the metric property itself is easy to show) of the form

$$d^{\leftrightarrow}_q(v, w) := [\,(d(v, w))^q + (d^{\leftarrow}(v, w))^q]^{1/q}\,. \tag{22}$$

Any of the preceding metrics (20), (21), or (22) lend themselves straightforwardly to various clustering techniques.

Example 4 *If d is the total variation or Hellinger distance, then the metric (20) takes values in the unit interval. It is then reasonable to automatically select a cutoff for a hierarchical clustering technique along the same lines as described in "Data reduction and anomaly detection" section, possibly after some rescaling. Here, we will consider total variation distance and use single linkage clustering.*

Recall the DCN of nonstop UK domestic flights described in "Taint bounds" section, but now over the entire week of 18 October 2010. Because this DCN has a rather idiosyncratic bipartite structure, it is not particularly instructive to look at its local temporal behavior à la (5): many of the transitions will be between airports and flights or conversely, rather than between two airports. Therefore, we consider the DCN a day at a time to avoid "getting stranded on a plane". The results are shown in Figs 5, 6, 7, 8, 9, 10 and 11.

For Monday, some of the likeliest transitions are as follows:

- *The transitions from CEG and FZO to FZO arise from a sequence $CEG \to FZO \to CEG \to FZO$ of three chartered flights between these airports, which host Airbus facilities and have no other scheduled passenger flights.*
- *The transitions from GLO, OXF, and TRE to SOU arise as follows. The first flight departing OXF departs to GLO at 0815, arriving at 0830. The first flight departing GLO not earlier than 0830 departs to IOM at 1025, arriving at 1130. The first flight departing IOM not earlier than 1130 departs to GLA at 1210, arriving at 1300. Meanwhile, the only flight departing TRE arrives at GLA at 1300. Starting from GLA*

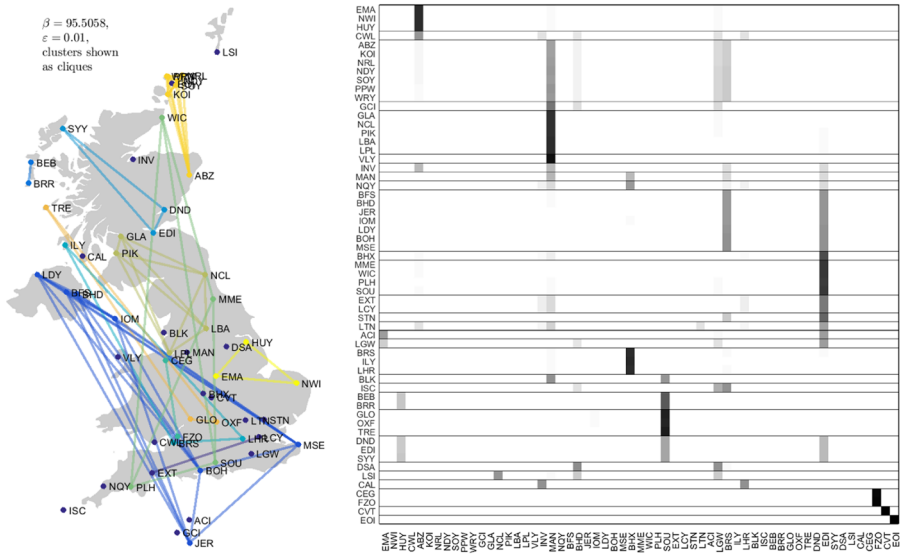

Fig. 5 Clusters from UK flight data for Monday, 18 October 2010 obtained using total variation distance in (20) and single linkage, with a dendrogram cutoff automatically selected along the lines mentioned in "Data reduction and anomaly detection" section. (L) Clusters displayed as cliques. (R) Transition matrix for $\beta = 1/\langle \Delta\tau \rangle$ and $\varepsilon = 10^{-2}$, with entries indicated by grayscale and indices/airports permuted to reflect clusters, which are themselves delineated by horizontal lines

at 1300 and taking the shortest possible (even instantaneous) layovers, we have the sequence of nominally connecting flights

$$GLA \underset{1315}{\overset{1425}{\rightarrow}} LTN \underset{1500}{\overset{1610}{\rightarrow}} GLA \underset{1610}{\overset{1730}{\rightarrow}} LCY \underset{1735}{\overset{1910}{\rightarrow}} EDI \underset{1920}{\overset{2020}{\rightarrow}} MAN \rightarrow SOU$$

that terminates at SOU when there are no more departing flights for the day.

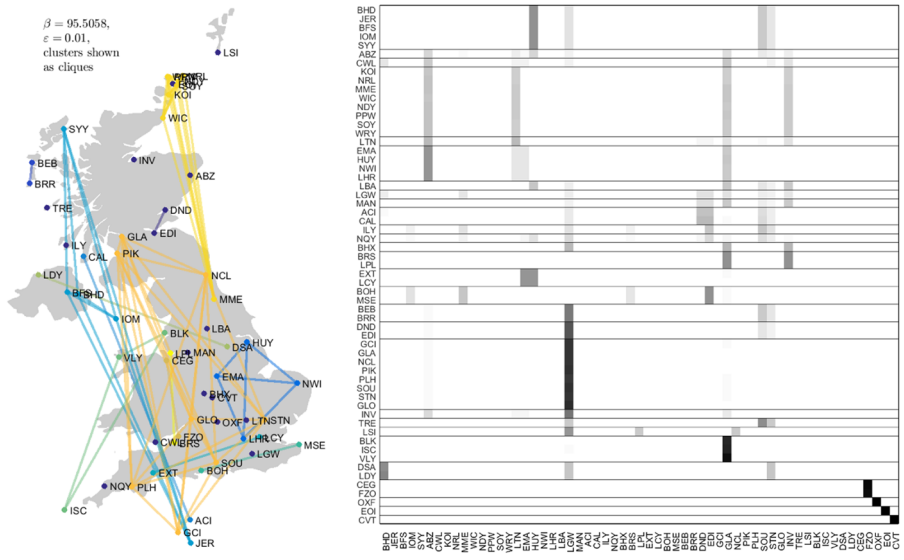

Fig. 6 Clusters from UK flight data for Tuesday, 19 October 2010; otherwise as in Fig. 5. Cluster colors only distinguish clusters from each other and do not suggest affinity across days

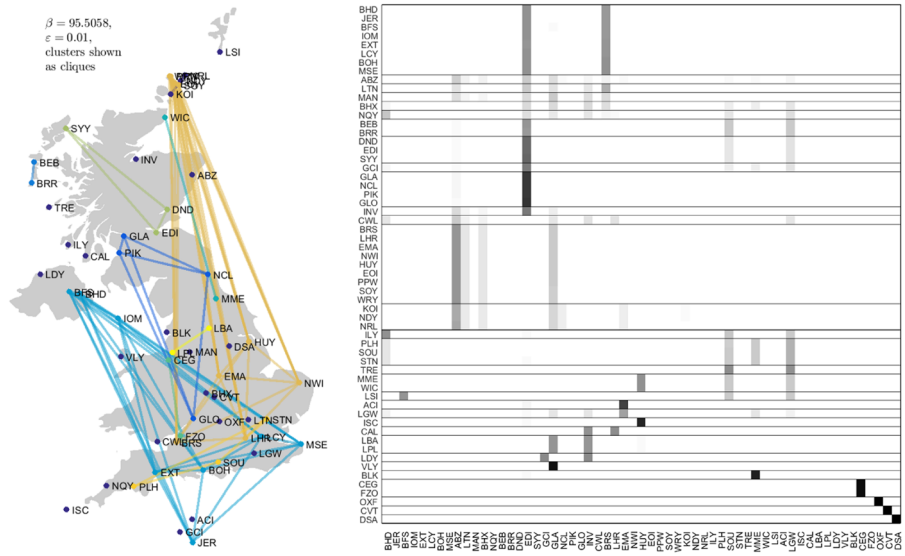

Fig. 7 Clusters from UK flight data for Wednesday, 20 October 2010; otherwise as in Figure 5. Cluster colors only distinguish clusters from each other and do not suggest affinity across days

- *The transitions from EMA, HUY, and NWI to ABZ arise as follows. The first flights departing EMA, HUY, and NWI each arrive at ABZ between 0800 and 0810; meanwhile, the first flight departing ABZ not earlier than 0800 departs to MME at 0820. Taking shortest possible layovers as above eventually terminates at ABZ when there are no more departing flights for the day.*

The "greedy" connections of the sort above become less likely as β decreases.

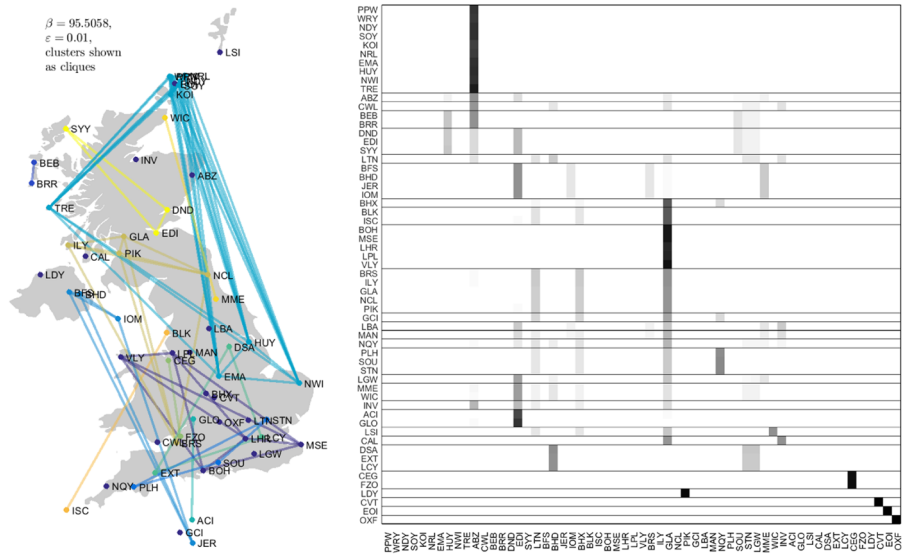

Fig. 8 Clusters from UK flight data for Thursday, 21 October 2010; otherwise as in Fig. 5. Cluster colors only distinguish clusters from each other and do not suggest affinity across days

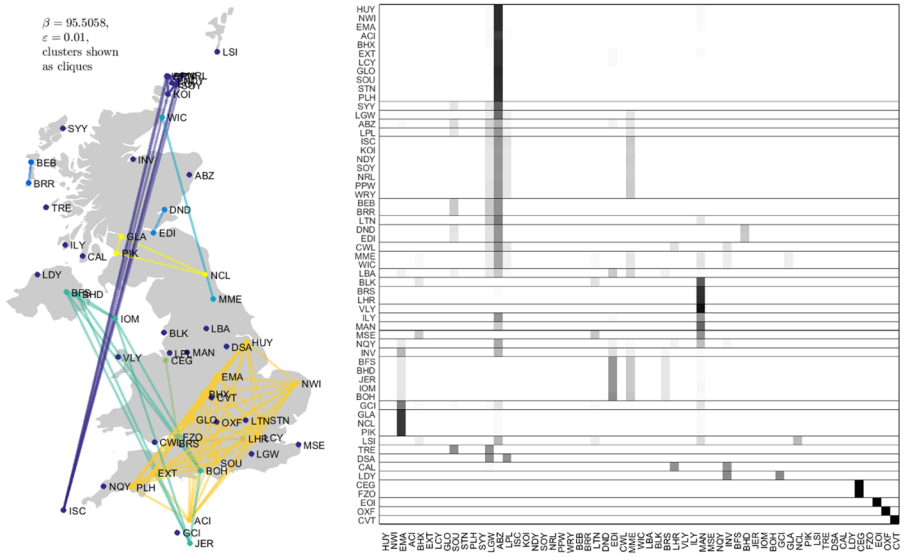

Fig. 9 Clusters from UK flight data for Friday, 22 October 2010; otherwise as in Fig. 5. Cluster colors only distinguish clusters from each other and do not suggest affinity across days

Remarkably, there is a considerable amount of geographic coherence to the clusters on weekdays, particularly on Friday. For example, even the two geographically largest clusters in the center panel of Fig. 5 consist entirely of airports on the periphery of the UK and Crown dependencies. As another example, the airports in Orkney with service (viz., KOI, NDY, NRL, PPW, SOY, and WRY) are part of a single, geographically local cluster. [9]

The time reversal (not shown) is less coherent, probabilistically and geographically. This highlights a critical distinction between DCNs with scheduled versus observed contacts. The former will generally correspond to something like a transportation network that is specifically engineered to facilitate certain connections (this in turn highlights that the UK domestic flight data is really a superposition of DCNs corresponding to individual airlines, with codeshares serving to add further complexity). It seems plausible that this distinction underlies the very limited degree of commonality between the forward and time-reversed clusters. □

Remarks

The model of "Markov chain models for DCNs" section exhibits a very sensitive dependence on the structure of the underlying DCN. In experiments not detailed here, we have seen that inserting just 1% of uniformly random contacts seriously degrades the model's data reduction and anomaly detection characteristics. Rather than being a shortcoming, this behavior demonstrates that the model actually captures delicate yet critical aspects of flows from contact data.

An interesting perspective on this model is that it yields a time-varying geometry once a metric on probability distributions is chosen. Using the induced metric to assess the

[9] As an amusing aside, the shortest scheduled passenger flight in the world is between PPW and WRY: it has been completed in under a minute.

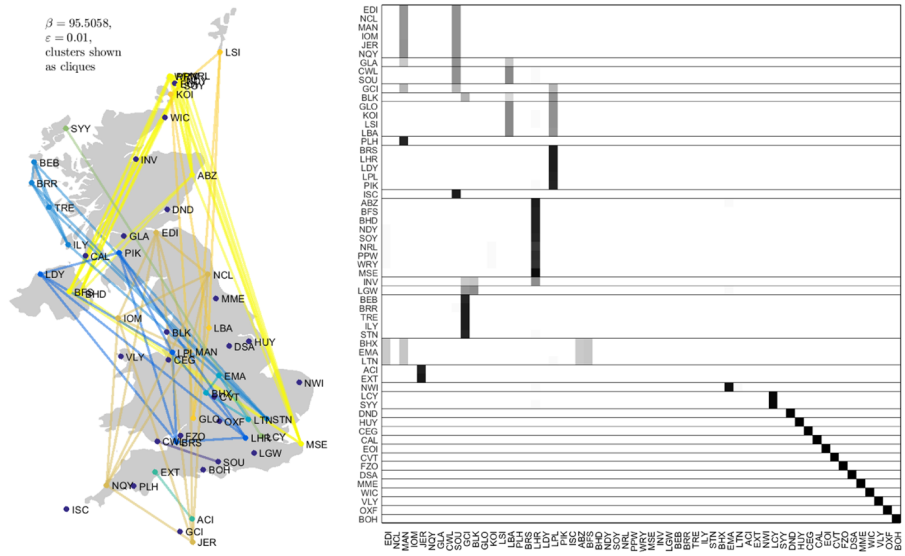

Fig. 10 Clusters from UK flight data for Saturday, 23 October 2010; otherwise as in Fig. 5. Cluster colors only distinguish clusters from each other and do not suggest affinity across days

model dynamics is worth examining in future work. In a similar vein, analyzing the variation of information (Meilă 2007) between subsequent clusterings would give additional principled insight into the behavior of DCNs.

Regarding the taint bounds of "Taint bounds" section, we note that using alternative arithmetics/semirings along similar lines is also of interest, such as the so-called log and Viterbi semirings.

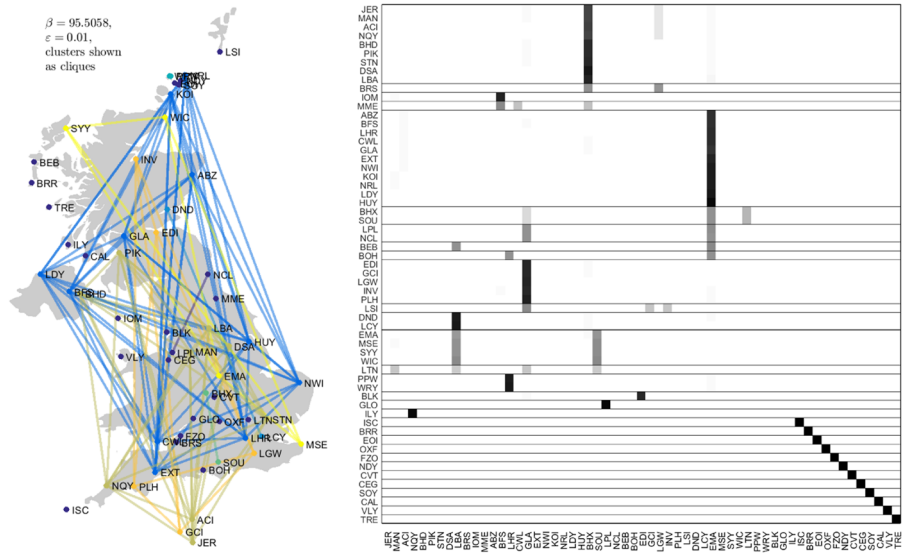

Fig. 11 Clusters from UK flight data for Sunday, 24 October 2010; otherwise as in Fig. 5. Cluster colors only distinguish clusters from each other and do not suggest affinity across days

Abbreviations
DAG: directed acyclic graph (In "Taint bounds" section and "Clustering" section we make extensive use of International Air Transport Association airport codes, for which see https://www.iata.org/services/Pages/codes.aspx.); DCN: directed contact network; FPR: false positive rate; NPV: negative predictive value; PPV: positive predictive value (or precision); RG: renormalization group; TPR: true positive rate (or recall); UK: United Kingdom

Acknowledgements
We thank Yingbo Song, Rob Ross, and Mike Weber for many helpful discussions as well as creating the summary and ground truth data used in "Data reduction and anomaly detection" section.

Authors' contributions
GC conceived of and prototyped the techniques discussed in "Taint bounds" section, provided advice and comments on the subject matter of the entire paper, and edited the manuscript. SH conceived of the model in "Markov chain models for DCNs" section, performed the data analyses presented in the paper, and wrote the manuscript. Both authors read and approved the final manuscript.

Funding
This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL). Cybenko's efforts in this work were also partially supported by ARO MURI Grant W911NF-13-1-042. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA, the US Army or AFRL.

Availability of data and material
Non-public datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Thayer School of Engineering, Dartmouth College, Hanover, NH 03755, USA. [2]BAE Systems FAST Labs, 4301 North Fairfax Drive, Arlington, VA 22203, VA, USA.

Received: 11 March 2019 Accepted: 10 September 2019
Published online: 15 November 2019

References
Bang-Jensen J, Gutin G (2009) Digraphs: Theory, Algorithms and Applications. 2nd. Springer, LOndon. https://doi.org/10.1007/978-1-84800-998-1
Barenblatt GI (2003) Scaling, Cambridge. https://doi.org/10.1017/cbo9780511814921
Barrat A, Cattuto C (2013) Temporal networks of face-to-face human interactions. In: Temporal Networks. Springer, Berlin. pp 191-216
Bassett DS, Porter MA, Wymbs, NF Grafton ST, Carlson JM, Mucha PJ (2013) Robust detection of dynamic community structure in networks. Chaos: Interdisc J Nonlinear Sci 23(1):013142
Bazzi M, Porter MA, Williams S, McDonald M, Fenn DJ, Howison SD (2016) Community detection in temporal multilayer networks, with an application to correlation networks. Multiscale Modeling Simul 14(1):1-41
Bianchi et al. FM (2016) Identifying user habits through data mining on call data records. Eng Appl Artif Intell 54:49-61
Brémaud P (1999) Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues. Springer, New York
Chan SC, et al. (2017) Expressiveness benchmarking for system-level provenance. TaPP
Chakrabarti D (2004) AutoPart: parameter-free graph partitioning and outlier detection. PKDD
Cheney J, Acar UA, Perera R (2013) Toward a theory of self-explaining computation. In: Tannen V, et al. (eds). Search of Elegance in the Theory and Practice of Computation. Springer
Flaška V, et al. (2007) Transitive closures of binary relations I. Acta Uni Carolinae - Math Phys 48:55
Gallotti R, Barthelemy M (2015) The multilayer temporal network of public transport in Great Britain. Sci Data 2:140056
Gallotti R, Barthelemy M (2015) The multilayer temporal network of public transport in Great Britain. Dryad Digit Repository. https://doi.org/10.5061/dryad.pc8m3
Gauvin L, Panisson A, Cattuto C (2014) Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. PloS One 9(1):e86028
Ge X, Parise S, Smyth P (2003) Clustering Markov states into equivalence classes using SVD and heuristic search algorithms. AISTATS
Glazek K (2002) Selected Applications of Semirings. In: A Guide to the Literature on Semirings and their Applications in Mathematics and Information Sciences. Springer. pp 67–87. https://doi.org/10.1007/978-94-015-9964-1_6
Goldenfeld N (1992) How Phase Transitions Occur in Principle. In: Lectures on Phase Transitions and the Renormalization Group. Addison-Wesley. pp 23–83. https://doi.org/10.1201/9780429493492
Grindrod P, Higham DJ (2013) A matrix iteration for dynamic network summaries. SIAM Rev 55:118
Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: first steps. Soc Netw 5:109
Holme P (2015) Modern temporal network theory: a colloquium. Eur Phys J B 88:234
Huntsman S (2018a) Topological mixture estimation. ICML
Huntsman S (2018b) A Markov model for inferring flows in directed contact networks. In: Aiello L, Cherifi C, Cherifi H, Lambiotte R, Lió P, Rocha L (eds). Complex Networks and Their Applications VII. COMPLEX NETWORKS 2018. Studies in Computational Intelligence. Springer, Cham Vol. 812

Jenkinson G, et al. (2017) Applying provenance in APT monitoring and analysis. TaPP
Karschau J, Zimmerling M, Friedrich BM (2018) Renormalization group theory for percolation in time-varying networks. Sci Rep 8(1):8011
King ST, Chen PM (2005) Backtracking intrusions. ACM Trans Comp Sys 23:51
Lambiotte R, Rosvall M, Scholtes I (2019) From networks to optimal higher-order models of complex systems. Nat Phys 15(4):313–320. https://doi.org/10.1038/s41567-019-0459-y
Lencastre P, et al. (2016) From empirical data to continuous Markov processes: a systematic approach. Phys Rev E 93:032135
Malliaros FD, Vazirgiannis M (2013) Clustering and community detection in directed networks: a survey. Phys Rep 533:95–142
Masuda N, Lambiotte R (2016) Models of temporal networks. In: A Guide to Temporal Networks. World Scientific. https://doi.org/10.1142/q0033
Meilă M (2007) Comparing clusterings-an information based distance. J Mutlivariate Anal 98:873
Newman ME, Watts DJ (1999) Renormalization group analysis of the small-world network model. Phys Lett A 263(4-6):341–346
Nielson F, Nielson HR, Hankin C (2010) Principles of Program Analysis. Springer, Berlin
Perra N, et al. (2012) Random walks and search in time-varying networks. Phys Rev Lett 109:238701
Ramsey NF (1956) Thermodynamics and statistical mechanics at negative absolute temperatures. Phys Rev 103:20
Rocha LEC, Masuda N (2014) Random walk centrality for temporal networks. New J Phys 16:063023
Rohe K, Qin T, Yu B (2016) Co-clustering directed graphs to discover asymmetries and directional communities. Proc Nat Acad Sci 113:12679
Rosvall M, Esquivel AV, Lancichinetti A, West JD, Lambiotte R (2014) Memory in network flows and its effects on spreading dynamics and community detection. Nat Commun 5:4630
Saramäki J, Holme P (2015) Exploring temporal networks with greedy walks. Eur Phys J B 88:334
Sarzynska M, Leicht EA, Chowell G, Porter MA (2015) Null models for community detection in spatially embedded, temporal networks. J Compl Netw 4(3):363–406
Schwartz EJ, Avgerinos T, Brumley D (2010) All you ever wanted to know about dynamic taint analysis and forward symbolic execution (but might have been afraid to ask). In: 2010 IEEE Symposium on Security and Privacy. https://doi.org/10.1109/sp.2010.26
Ser-Giacomi E, et al. (2015) Most probable paths in temporal weighted networks: an application to ocean transport. Phys Rev E 92:012818
Speidel L, Takaguchi T, Masuda N (2015) Community detection in directed acyclic graphs. Eur Phys J B 88:203
Starnini M, et al. (2012) Random walks on temporal networks. Phys Rev E 85:056115
Valdano E, Poletto C, Colizza V (2015) Infection propagator approach to compute epidemic thresholds on temporal networks: impact of immunity and of limited temporal resolution. Eur Phys J B 88:341
Valdano E, Fiorentin MR, Poletto C, Colizza V (2018) Epidemic threshold in continuous-time evolving networks. Phys Rev Lett 120(6):068302

Publisher's Note