Applied Network Science

RESEARCH Open Access

Network memory in the movement of hospital patients carrying antimicrobial-resistant bacteria

Ashleigh C. Myall[1,2], Robert L. Peach[1,5], Andrea Y. Weiße[2,3], Siddharth Mookerjee[4], Frances Davies[4], Alison Holmes[4] and Mauricio Barahona[1*]

*Correspondence:
m.barahona@imperial.ac.uk
[1] Department of Mathematics, Imperial College London, London, UK
Full list of author information is available at the end of the article

Abstract

Hospitals constitute highly interconnected systems that bring into contact an abundance of infectious pathogens and susceptible individuals, thus making infection outbreaks both common and challenging. In recent years, there has been a sharp incidence of antimicrobial-resistance amongst healthcare-associated infections, a situation now considered endemic in many countries. Here we present network-based analyses of a data set capturing the movement of patients harbouring antibiotic-resistant bacteria across three large London hospitals. We show that there are substantial memory effects in the movement of hospital patients colonised with antibiotic-resistant bacteria. Such memory effects break first-order Markovian transitive assumptions and substantially alter the conclusions from the analysis, specifically on node rankings and the evolution of diffusive processes. We capture variable length memory effects by constructing a lumped-state memory network, which we then use to identify individually import wards and overlapping communities of wards. We find these wards align closely to known hotspots of transmission and commonly followed pathways patients. Our framework provides a means to focus infection control efforts and cohort outbreaks of healthcare-associated infections.

Keywords: Memory networks, Patient pathways, Mobility patterns, Healthcare networks, Infectious disease, Antimicrobial-resistance

Introduction

Antimicrobial resistance (AMR) poses one of the greatest risks to human health (Prestinaci et al. 2015). Currently, around 700,000 people worldwide die from infections with resistant pathogens each year, and this number is estimated to rise to up to 10 Million by 2050 (Interagency Coordination Group on Antimicrobial Resistance 2019; O'Neill 2016). Hospitals and other healthcare facilities act as key vectors for the spread of AMR through healthcare-associated infections (HAI) (Struelens 1998). Persistent colonisation of hospital patients and the networked nature of hospital processes underlying patient mobility will likely cause AMR to remain prevalent (Pastor-Satorras and Vespignani 2001b). Several factors moreover exacerbate the spread of AMR in healthcare facilities, including the selective pressures generated by increased antimicrobial usage, and the large pool of vulnerable patients, who are more susceptible to infections (Organization 2002). The need for infection prevention and control (IPC) can therefore not be understated.

Understanding the transmission dynamics of AMR promises valuable insights to improve IPC strategies. Key to these measures will be the analysis of patient pathways capturing the movement of patients carrying AMR during their hospital stay. Like many real-world systems, healthcare facilities have complex structure, which when ignored can limit the insights into the underlying dynamic processes. In this study we focus on mapping the movement pathways of patients known to carry antimicrobial-resistant bacteria onto physical structures of the hospitals. Specifically, we focus on patients colonised with Carbapenemase-producing *Enterobacteriaceae* (CPE). CPE is a particularly concerning form of AMR that confers resistance to carbapenems, broad-spectrum antibacterials often used as last-line antibiotics. CPE infections have recently seen a global surge amongst HAIs (Bonomo et al. 2018; Logan and Weinstein 2017).

Networks provide a powerful formalism to analyse the movement of patients in hospitals. Nodes typically represent physical locations within the hospital, such as wards, and edges represent the flow of patients between these locations, with edge weights encoding the volume of patient flow from one location to another. To facilitate analysis, we can aggregate the movements of individual patients into probabilities of transitioning between hospital wards (Donker et al. 2012; Bean et al. 2017). Typically, patient trajectories are broken down into individual transitions between wards: first, the number of transitions between each ward is summed across all patients and subsequently, for each ward the sum of all out-going transitions is normalised to one. The constructed network may then be interpreted as a first-order Markov model, where a random walker transitions with a probability proportional to the observed outflow volume from the current node to others in the network (Salnikov et al. 2016).

This dynamical assumption, whilst useful because of its simplicity and ease of implementation, is however limited by the assumption that transitions between nodes are independent of prior nodes within the patient pathway. Previous studies have indeed shown that first-order Markovian dynamics are not sufficient to fully model network dynamics of disease propagation (May and Lloyd 2001; Pastor-Satorras and Vespignani 2001a). Akin to shipping trajectories or passenger movement between airports, patient movement in hospitals tends to follow particular patterns dictated by medical or operational constraints. In particular, it is plausible that patient trajectories could bear 'memory', that is, a subsequent move depends on several or all previous locations visited, and not solely on the current location leading to transitive dependence in the data.

Introduced by Shannon (1948), higher-order memory models have shown relevance across a number of applications, and a wide range of real world movement data (Song et al. 2010; Kareiva and Shigesada 1983; Chierichetti et al. 2012; Singer et al. 2014; Gonzalez et al. 2008; Heath et al. 2008) including several epidemiological data sets (Balcan and Vespignani 2011; Poletto et al. 2013). Ignoring such transitive dependencies and modelling patient movement via memory-less, first-order Markov models can distort both network topology and conclusions on the underlying process (Mucha et al. 2010). Despite the clear importance of transitive dependence, to date we only find one study

(Palla et al. 2017) of hospital patient movement accounting for these relationships, and none when looking at AMR across healthcare facilities. Hence, in this study we investigate evidence for and implications of *transitive dependencies* in the movement patterns of hospital patients colonised with a CPE by including *memory* in our network models.

To model these effects, we use *memory networks*, which encode the memory of individual trajectories into higher-order transitive relationships, and which have successfully been used to investigate transitive dependence in pathway data (Lambiotte et al. 2019). To provide some intuition behind memory networks, consider a simple example of a small network of a hospital with five wards where the patients can follow one of two possible routes between the wards, and the two routes share one common node (Fig. 1a). A first-order (memory-less) network model assuming full transitivity (Fig. 1b) would wrongly suggest that a patient starting at v_1 could transition to v_5 with some probability, when in fact, only patients starting at v_2 can reach v_5. In a memory network (Fig. 1c) these transitive dependencies are captured by abstracting away from a network of physical nodes to a higher-order networks of state nodes representing the possible dynamical states of the system (i.e., the sequence of hospital wards visited up to a given memory) (Edler and Bohlin 2017). For a memory network, the order k, determines the number of previous pathway steps to consider in the state network, acting to parameterise 'memory'. Such memory is

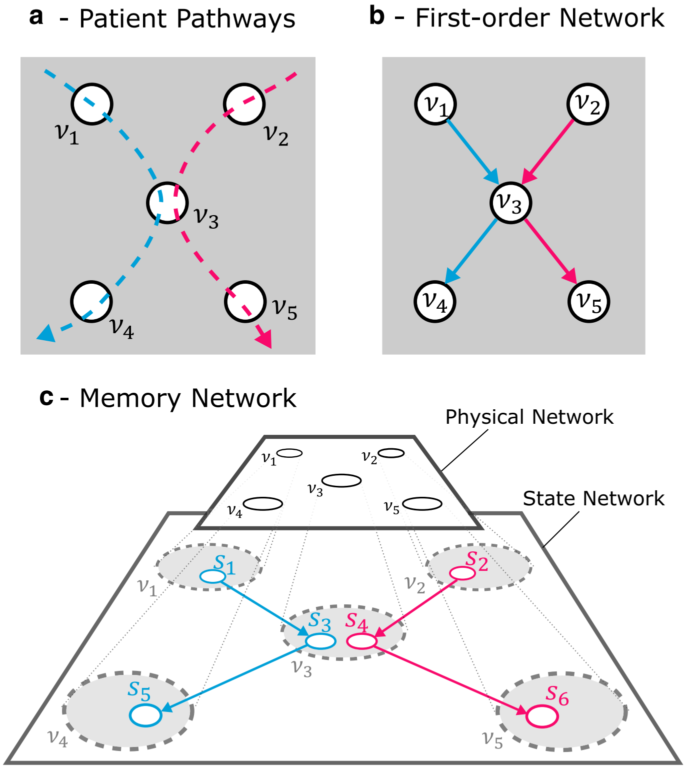

Fig. 1 Illustration of transitive dependence encoded into memory network. **a** Two sets of typical patient pathways, largely independent, but passing through the same ward as an intermediate point in their pathways. **b** First-order representation of **a** without any memory (**c**) Memory network representation **a**, whereby a physical node network maps to state nodes, which encode transitive dependence of the patient pathways and constrain a random walkers movement

directly incorporated into the state network though each state node representing a pathway of length $k-1$ present in the trajectories. This state network can be thought of as an additional layer of information still bound to the physical network since each state node is assigned to a physical node. The state network thus acts to constrain how a random walker transitions between physical nodes. These higher-order network abstractions lend themselves to learning tasks that can pinpoint key properties underlying the dynamical process. In the case of HAIs, this can offer insight into more accurate patterns in the movement of infected patients otherwise lost in a network model that assumes full transitivity.

Below we present the analysis of patients pathways confirmed to be colonised with CPE. We begin by presenting our data and a description of the hospital network. We then present evidence for memory within patient pathways by contrasting models constructed with and without memory. Next, we construct a lumped state memory network, which captures transitive dependence and removes redundancy. We carry out multiscale community detection on this network, and present the resultant communities, highlighting specific wards and specialities that are important across different regions of the network. Finally, from a clinical point of view, we discuss how these results can aid infection prevention and control to identify hospital structures that are relevant for disease transmission and thus to focus intervention strategies.

Results

Data

Our analysis is based on anonymised electronic health records of patients from a large 1000-bed Trust of London teaching hospitals. Specifically, we used ward-level movement patterns of 967 patients who tested positive for CPE over a period of two years between 2018 and 2020. We focused on the subset of 526 patients who moved between at least two wards during their hospital stay for a total of 1958 transitions between 96 hospital wards.

Formally, the hospital Trust is structured around 17 *specialities* and 19 *buildings*, the latter belonging to three *hospital sites* (Fig. 2). Hospital site 3 acts as a Tertiary site with only speciality wards. Whilst sites and buildings are constrained by geographical factors, specialities are defined by medical procedures and thus may overlap across sites and buildings. In fact, a number of specialities span all five hospital sites (Critical Care, Elderly Care, Medicine, Private, and Surgery). Geographical structures constrain patient movement to some extent: patients with certain co-morbidities and therapeutic requirements are likely to be constrained to a single or several specialities supporting those needs, whereas other patients can move within buildings, or between wards placed closely for logistics and ease of transfer.

From patient pathways to network models

We consider the trajectories of p patients. Each patient pathway as a trajectory T_α and the set of $\alpha = 1, \ldots, p$ trajectories is $\mathcal{T} = \{T_1, T_2, T_3, \ldots, T_p\}$. Each T_α consists of a time-ordered set representing discrete movements between nodes,

$$T_\alpha = \{v_i \rightarrow v_j \rightarrow \ldots \rightarrow v_k\}\,, \tag{1}$$

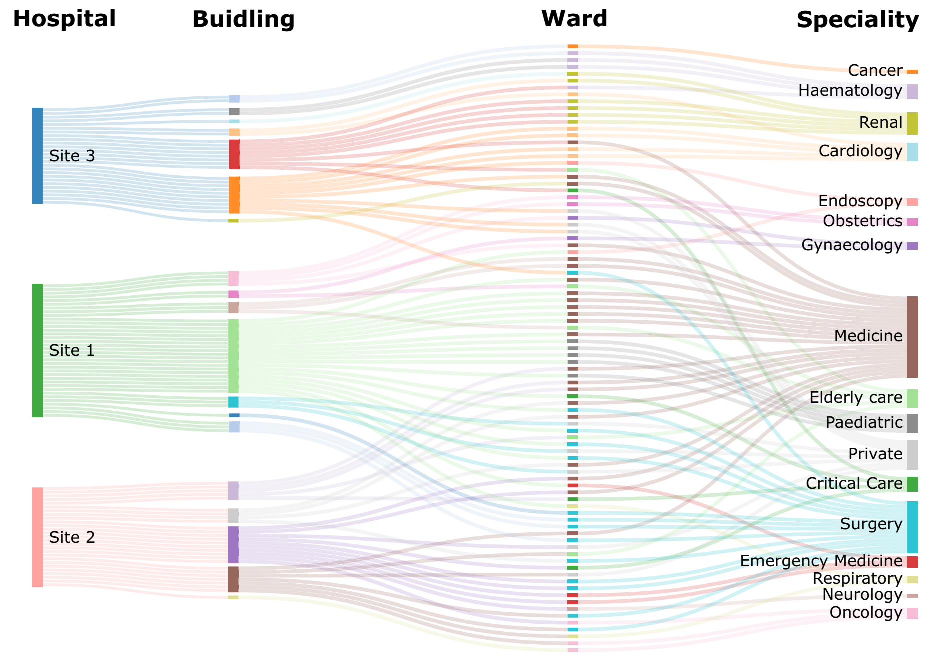

Fig. 2 Sankey diagram of Trust Structure traversed by CPE patients. Broken down by hospital site, and buildings to wards, then also broken down by speciality into wards

where each node refers to one of N hospital wards $\mathcal{N} = \{v_1, v_2, v_3, \ldots, v_N\}$. Since these nodes represent physical locations, we will refer to them as *physical nodes* to avoid confusion with *state nodes*, which we introduce next.

In order to understand the aggregate dynamics of all patients, whilst preserving memory effects in $\mathcal{T}$, we represent the trajectories as a memory network as proposed by Rosvall et al. (2014). This way, we maintain information about physical nodes $\mathcal{N}$ whilst instilling transitive dependence in the connectivity patterns of an underlying state-network, $\mathcal{M}_k = (\mathcal{E}_k, \mathcal{S}_k)$. Here $\mathcal{E}_k$ is the set of edges that link the set of state nodes $\mathcal{S}_k$ that capture higher-order memory of order k (Edler and Bohlin 2017).

A memory network of order $k = 1$, $\mathcal{M}_1$, represents a system with zero memory, where the movement of a random walker only depends on its current location. In this special case, the state network $\mathcal{M}_1 = (\mathcal{E}_1, \mathcal{S}_1)$ is equivalent to an aggregated physical network, and the set of states directly maps to the set of physical nodes, i.e., $\mathcal{S}_1 = \mathcal{N}$. The edge weights w_{ij} conforming the set $\mathcal{E}$ in $\mathcal{M}_1$ represent the frequency of transitions between physical nodes v_i and v_j across the set of trajectories $\mathcal{T}$. Given w_{ij}, we can write the transition probability matrix P_1 for $\mathcal{M}_1$ as

$$p_{ij} = P(i \to j) = \frac{w_{ij}}{\sum_j w_{ij}}. \tag{2}$$

In memory networks of higher-order, where $k > 1$, state nodes represent pathways of length $k - 1$, and are no longer equivalent to the physical nodes $\mathcal{S}_k \neq \mathcal{N}$. This representation allows us to introduce the memory dependence in $\mathcal{T}$, capturing multi-step patterns of flow via the state nodes of the network (Salnikov et al. 2016).

In particular, for the second-order memory network $\mathcal{M}_2$, a state node represents a directed pathway of length one $s_j = \overrightarrow{ij}$. For two states nodes $s_j = \overrightarrow{ij}$ and $s_\ell = \overrightarrow{j\ell}$ to be connected, a path of length two, $(v_i \rightarrow v_j \rightarrow v_\ell)$, must occur in the set of trajectories $\mathcal{T}$. Similarly for higher-order models, edges between state nodes are weighted $w_{s_j s_l}$ and capture the number of occurrences that a transition between state nodes s_j and s_l was observed in $\mathcal{T}$. Transition probabilities P_k of $\mathcal{M}_k$ for any order can be derived from Equation 3 by altering the state node set S to represent pathways of length $k-1$, so

$$p_{s_i s_j} = P(s_i \rightarrow s_j) = \frac{s_{ij}}{\sum_j s_{ij}}. \tag{3}$$

Each state node can be mapped to a physical node (Fig. 1a), using an $|\mathcal{S}_k| \times N$ indicator matrix D, the elements of which, $D_{sv} \in \{0, 1\}$, indicate the final physical node of a pathway s.

We first constructed a first-order memory network $\mathcal{M}_1$ that contains 96 state nodes with a one-to-one mapping to the 96 wards (physical nodes). $\mathcal{M}_1$ consists of four weakly connected components, one of which contains the majority of state nodes (87 out of 96) (Additional file 1: Fig. S1). We next constructed a second-order memory network $\mathcal{M}_2$ that contains 384 state nodes, in 18 weakly connected components. Similarly, $\mathcal{M}_2$ consists of a single weakly component that contains the majority of state nodes (329 of 384). Structurally, $\mathcal{M}_1$ has a higher connectivity with a clustering coefficient of 0.287 and a diameter of 6, whereas $\mathcal{M}_2$ is more sparse with a clustering coefficient of 0.003 and a larger diameter of 31, resembling a series of connected line graphs (Additional file 1: Fig. S1).

Patient trajectories break first-order dynamics

Using random walks to reveal and probe the structure of networks has long been a foundational tool in network science (Masuda et al. 2017). A random walk is a stochastic process which consists of a succession of random steps with no memory of its past locations; however, in a system where transitive dependence plays a important role, a purely random walk becomes inaccurate and potentially misleading. Higher-order memory networks can capture deviations from first-order transitive assumptions by constraining where a random walker can next move depending on its previous location(s). For pathway models without transitive dependence, a random walker should be no more constrained when moving from a first-order memory network, $\mathcal{M}_1$, to a second-order memory network $\mathcal{M}_2$ (Rosvall et al. 2014). However, pathways exhibiting transitive dependence will constrain a random walker comparatively more in the second-order memory network. Here, we use the entropy rate of the random-walk to measure the uncertainty of moving between two state nodes (Shannon 1948; Schaub et al. 2012):

$$H(S_{t+1}|S_t) = \sum_{i,j} \pi(i) p(i \rightarrow j) \log p(i \rightarrow j), \tag{4}$$

where π denotes the stationary distribution across $\mathcal{M}$, and $p(i \rightarrow j)$ are the transition probabilities. A higher entropy rate reflects a higher uncertainty in the transitions of a random walker (Schaub et al. 2012). If a large amount of memory is important, we expect a decrease in the uncertainty of the random walker when accounting for the higher-order effects, since a random walker becomes more constrained. On the other hand, if memory plays little role, and does not constrain the transitions of the random walker, we would expect little change in entropy when accounting for memory.

We constructed memory networks $\mathcal{M}_k$ of order $k = 1, 2, 3, 4$ (description of the number of state nodes, edges and pathways for each $\mathcal{M}_k$ is detailed in Fig. 3a). Computing the entropy for each $\mathcal{M}_k$ we find increasing restriction of the random walk (reduced entropy) for larger k (Fig. 3b). In particular, we observe a large decrease in entropy from 2.70 to 0.57 when we move from first-order memory to second-order memory. Patient pathways with little to no memory effect would not exhibit any large change in entropy when moving from $\mathcal{M}_1$ to $\mathcal{M}_2$ and thus our results suggest that there exist patient pathways which break first-order Markovian transitive assumptions and highlight the importance of capturing memory.

Now we must determine the optimal order k for a given analysis. For small data sets, it is difficult to statistically validate whether memory networks with higher-order are relevant, given that the parameter space and complexity increases exponentially (Scholtes 2017). A common workaround is to use cross-validation, a model validation technique borrowed from machine learning (Singer et al. 2014). In cross-validation, data is partitioned and performance is determined as an average across partitions to reduce overfitting and selection bias (Arlot and Celisse 2010; Cawley and Talbot 2010). To perform cross-validation in the framework of a memory network we compute the rank orders of wards using a training set of patient pathways and then compare with the rank order of wards generated from visitation probabilities of a withheld partition of patient pathways. Similar to Rosvall et al. (2014), we used a generalised PageRank for higher-order models where the visitation probabilities of state nodes were summed for each physical node (see methods section: "Higher-order PageRank" section). The rank orders between train

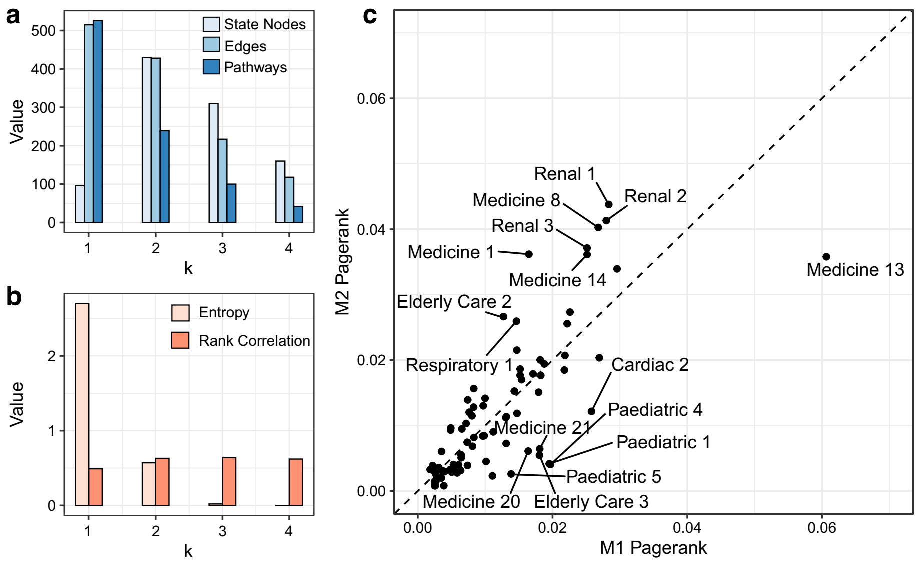

Fig. 3 A comparison of the effect of memory for models of order k. (**a**) Details of network size (no. state nodes, edges) and data (no. trajectories) for $k = 1, 2, 3, 4$. (**b**) The entropy of each higher order memory network and the rank correlation between train and validation sets. (**c**) PageRank of wards for $\mathcal{M}_1$ and $\mathcal{M}_2$, dashed line indicating equality of PageRank

and test sets were compared with Kendall–Tau rank correlation (Kendall 1938) and the results were averaged over a fivefold cross-validation split. We found that $\mathcal{M}_2$ was more predictive of the node ranking of physical nodes than $\mathcal{M}_1$ (0.60 to 0.49) (Fig. 3b). Moreover, across all folds the ranked correlation was more significant in $\mathcal{M}_2$ than $\mathcal{M}_1$ (Additional file 1: Table S1). This increased performance in $\mathcal{M}_2$ again suggests that a patient's current and previous location both affect future movement, and that accounting for this memory effect yields more accurate approximations of patient movement.

Whilst further higher-order memory effects may exist, we were unable to detect any increased predictive power beyond the second-order (Fig. 3b). We note that this may be due to limitations of our data; as we increase the order k, we must discard additional patient pathways with fewer than k transitions between wards. This is evident in Fig. 3a that shows a decreasing number of pathways as we increase the order; and whilst the number of state nodes and edges initially increases from the first-order to the second-order, as you may expect by increasing model complexity, due to the decreasing number of pathways we instead observe a decrease in the number of state nodes beyond the second-order. Herein, to retain enough patient pathways for reliable insights, we thus shall focus on the second-order memory network.

We then compared the PageRank of physical nodes (wards) between $\mathcal{M}_1$ and $\mathcal{M}_2$ (Fig. 3c). Whilst we found the PageRank of wards in $\mathcal{M}_1$ and $\mathcal{M}_2$ to be correlated ($\sigma = 0.81$, p val<0.01), there were a number of key deviations. In particular, we find three renal wards (Renal 1, 2 & 3) with a relatively higher ranking in $\mathcal{M}_2$, indicating that CPE patients frequently visit these wards. These results are reassuring, given that patients undergoing renal therapies at our institution were previously noted to have an increased risk of CPE acquisition (Otter et al. 2020), though whether this is unique to our institution or a more specific trait of this patient group is to be determined. The *higher* ranking of these Renal wards in $\mathcal{M}_2$ highlights the importance of using a constrained state node network to understand the clinical movement of these patients. Conversely, Medicine 13 was the highest ranked ward in $\mathcal{M}_1$, but was found to have a relatively lower rank in $\mathcal{M}_2$. Medicine 13 is an acute medical admissions unit, and as such acts as the entry/re-entry point for many patients to the hospital, rather than a transition ward or a ward which offers care, and whilst it plays a starting role in many patient pathways, it is seldom observed at any other point in a patients trajectory through the hospital.

Investigating memory effects with a discrete diffusive process

One way we can study the effect of memory is through the direct observation of its influence on a diffusion process starting at various points in the network (Lambiotte et al. 2019). Figure 4a, b displays the evolution of a discrete-time diffusive process for $\mathcal{M}_1$ and $\mathcal{M}_2$, each encoded by their respective transition matrix P_k, when injecting an impulse at a single ward (Medicine 13). At time $t = 0$, the diffusive process is entirely contained within the state node(s) corresponding to Medicine 13 (beyond the first-order, where physical nodes can have several state nodes, we share initial probability over states based on the frequency of pathways in $\mathcal{T}$ that $s_j = \overrightarrow{ij}$ represented). For times $t > 0$ we compute the probabilities of being on a given ward at time t through powers of the transition matrix P_k^t.

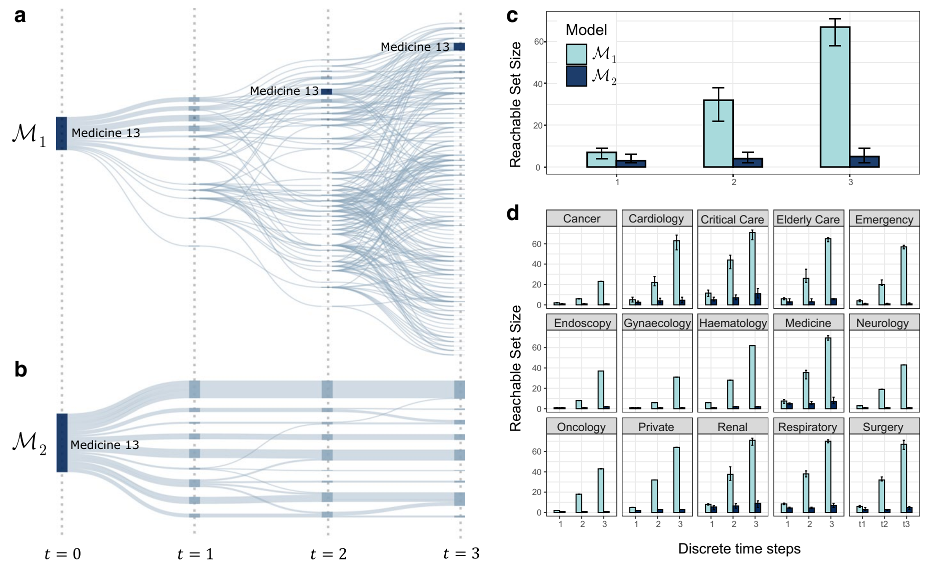

Fig. 4 Memory effects on network pathways. **a**, **b** A discrete random walk starting at Medicine 13 in the state networks of $\mathcal{M}_1$ and $\mathcal{M}_2$, modelled as a discrete time diffusion over the transition matrices P_1 and P_2 respectively. After a single discrete time step $t = 1$ the reachable number of wards is similar between $\mathcal{M}_1$ and $\mathcal{M}_2$, however, they quickly diverge as the diffusion evolves over time and the transitive effects increase in $\mathcal{M}_1$. **c**, **d** Median size of the reachable set of wards for $\mathcal{M}_1$ and $\mathcal{M}_2$: **c** Overall reachability after *t*-discrete steps, and **d** reachability after *t*-discrete steps broken down by speciality (in the largest connected component)

After a single discrete step $t = 1$ we find there is little effect of memory with the total number of wards reachable being similar for $\mathcal{M}_1$ and $\mathcal{M}_2$ (12 wards vs 9 wards, respectively). However, as we extend the diffusive process to $t = 2$ and $t = 3$ we find that the number of reachable wards from Medicine 13 increases rapidly for $\mathcal{M}_1$ (36 wards at $t = 2$, then 71 wards at $t = 3$) whereas we do not see any change in $\mathcal{M}_2$ (9 wards at $t = 2$, and 9 wards at $t = 3$). In fact, for $\mathcal{M}_1$ a random walk initialised at Medicine 13 can reach 71 out of the 79 wards within the largest weakly connected component in $\mathcal{T}$ after just 3 steps. This level of transitivity is not present in $\mathcal{T}$, and its absence is directly observable by looking at the restriction of flow evident in $\mathcal{M}_2$ (Fig. 4a, b). This difference comes from patients not starting at Medicine 13, but passing through its neighbours influencing the 2-step network transitivity.

Interestingly, $\mathcal{M}_2$ constrained walkers such that no backtracking to Medicine 13 is possible over the first 3 discrete transitions, in contrast to $\mathcal{M}_1$, where backtracking to Medicine 13 is possible for $t > 1$. In fact, using $\mathcal{M}_1$ there is a relatively large probability to revisit Medicine 13 after 2 or 3 steps ($p^2_{med13} = 0.18$ and $p^3_{med13} = 0.24$). Given that Medicine 13 is commonly an entry point/readmission point where patients go when waiting for diagnosis, we would expect a minimal backtracking effect in patient movement across short time frames since they move into subsequent specialities for treatment once a diagnosis is known. Hence, including memory through $\mathcal{M}_2$ better captures true patient flow.

Forward reachability is varyingly constrained by memory

We expanded the above framework to examine reachability across the entire network by performing the analysis for every possible starting node. For each ward, we compute the set of reachable wards after t time-steps and in Fig. 4c we display the median size of reachable sets for all wards under $\mathcal{M}_1$ and $\mathcal{M}_2$. Similar to the analysis of Medicine 13 in Fig. 4a, b, we find that the median size of reachable sets is relatively similar between $\mathcal{M}_1$ and $\mathcal{M}_2$ at $t = 1$. However, as t increases we again observe divergence in the reachable set sizes due to the significantly larger set of reachable wards in the first-order model $\mathcal{M}_1$. Indeed, after 3 time-steps only 5 wards are reachable on average under $\mathcal{M}_2$ as compared to the 79 reachable wards under $\mathcal{M}_1$. Hence $\mathcal{M}_1$ is inflating transitivity between wards and distorting the set of reachable wards for a patient through inherent ignorance of prior ward visits. We also observe that the variance of the reachable set of wards for $\mathcal{M}_1$ increases for $t = 2, 3$, suggesting that the importance of memory is different depending on the ward at which the diffusive process was initialised.

To study this, we next break down wards by speciality and examine the importance of memory on the median set size of reachable wards. Figure 4d summarises the size of reachable sets averaged across wards within the same speciality. We notice that specialities which are known to be well visited by CPE patients in this hospital setting (e.g., Critical Care, Renal) exhibit a comparatively larger reachability set size when compared to the aggregated view in Fig. 4c. In contrast, specialities such as Neurology or Cancer which are less common to CPE patients exhibit a relatively lower reachability. These different reachabilities between specialities likely result from two mechanisms: (1) the different roles specialities play within the network and their transitivity by CPE patient trajectories, and (2), that memory effects may vary in different areas of the network, i.e. the extent to which a previous ward determines a patients next move. Hence, it may be optimal to construct a 'hybrid' of $\mathcal{M}_1$ and $\mathcal{M}_2$ which incorporates many of the desirable memory effects in $\mathcal{M}_2$, but simplifies parts of the model where greater transitivity is in fact present.

Reducing complexity using state node lumping

Given a large set of trajectories, the problem arises that state node networks $\mathcal{M}_k$ can become increasingly large and often duplicate or contain redundant information. In the case of patient trajectories, not all hospital pathways may exhibit memory effects in equal measure. Variable-length Markov models, pioneered by Rissanen (1983) alleviate some of these issues by introducing a 'lumping' step in which 'redundant' states are merged, thus enabling models to capture variable lengths of memory and remove model redundancy (Jääskinen et al. 1983; Bühlmann and Wyner 1999). Remembering that in memory networks, state nodes are assigned to physical nodes, we will often find several state nodes that are connecting the same physical nodes just via different edges. There is no need for this repetition and therefore here we focused on lumping state nodes within the same physical node to form so called 'meta-state nodes' or 'lumped nodes' which also benefit from preserving the physical network structure (Lambiotte et al. 2019). For each lumped node, we reassemble all connections between two states nodes such that weighting and connectivity are preserved (Edler and Bohlin 2017). In effect, 'lumping' nodes retains *relevant* and distinct patterns of transitive dependence in the original pathways;

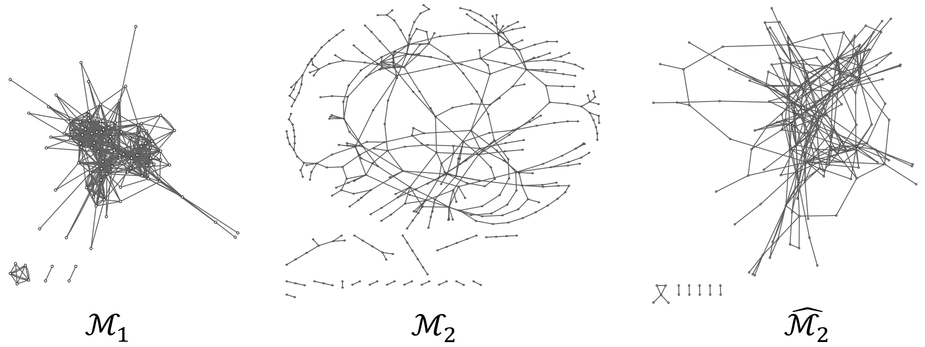

Fig. 5 From first-order network to second-order network and everything between. The first-order network $\mathcal{M}_1$ (left), the lumped second-order network $\hat{\mathcal{M}}_2$ (middle), and the second-order state node network $\mathcal{M}_2$ (right) ordered by scale of model complexity

however, for our purposes it also serves to 'de-sparsify' $\mathcal{M}_2$, improving its practicality and making it useful for subsequent learning tasks that assume greater connectivity.

In our approach, we lump together state nodes based on the similarity of visitation probabilities computed from a discrete diffusive process encoded in the state transition matrix P_k over t-steps. Existing node lumping methodologies use a 1-step random walk to identify state nodes that have similar connectivity within the network (Edler and Bohlin 2017; Persson et al. 2016). Here, however, we extend this approach to t-steps to identify similarity across a greater network locality. Using an agglomerative clustering method on the discrete diffusive process, we can then identify state nodes with similar connectivity, and if both are members of the same physical node they can be lumped together (Hastie et al. 2009). Importantly, this agglomerative clustering is parameterised by a clustering rate r, which controls the total proportion of state nodes to lump (for a detailed explanation refer to methods section: State lumping on local connectivity).

To what extent should we lump state nodes together? At one extreme, we have the state node network $\mathcal{M}_k$ without any lumping and at the other extreme we have the physical node network where every state node has been lumped together within its respective physical node. We want to identify an optimal amount of lumping, comfortably between the two extremes, that retains transitive dependence but removes redundant or duplicated information. The resulting lumped network is denoted $\hat{\mathcal{M}}_k$. In order to quantitatively determine the optimal lumping, we used 'ground-truth' community structures such as buildings, specialities, and hospital sites and compared these annotations with the results of community detection on the lumped network $\hat{\mathcal{M}}_k$. Whilst these structures do not fully constrain patient movement and therefore cannot provide an exact ground truth, there does exist a correlation with patient movement. We hypothesised that the optimal lumping would be found at the elbow of a fitness curve generated from the ability to detect known hospital structures in community structures, thus providing a trade-off between model accuracy and simplicity. Accordingly, we found that a clustering rate of $r = 0.35$ gave the optimal lumped model (Additional file 1: Fig. S3).

The lumped network $\hat{\mathcal{M}}_2$ contains 171 state nodes across 7 weakly connected components. Similar to the state node networks $\mathcal{M}_1$ and $\mathcal{M}_2$, we found a large weakly connected component that contained the majority of state nodes (156 out of 171) (Fig. 5).

Aside from visually appearing to exist in a state between $\mathcal{M}_1$ and $\mathcal{M}_2$, both its clustering coefficient (0.054) and network diameter (11) sat comfortably between $\mathcal{M}_1$ and $\mathcal{M}_2$, serving to validate its balance of complexity, connectivity, and higher-order dependencies. Note that unlike $\mathcal{M}_2$, the lumped state network $\hat{\mathcal{M}}_2$ no longer resembles a series of lines graphs, and hence provides a more practical structure over which to apply community detection.

Community detection reveals overlapping clusters of wards common to distinct pathways

By constraining a walkers movement within the connectivity patterns of $\mathcal{M}_k$, for $k > 1$, we can identify communities within $\mathcal{M}_k$ that conserve flow from a dynamical perspective. Given that $\mathcal{M}_k$ is composed of state nodes, the memory-dependent structure $\mathcal{C}$ will provide network partitions that shed light into community structure. Here we use Markov Stability (MS), a quasi-hierarchical community detection algorithm that identifies regions within a network in which a diffusive process becomes transiently constrained (Delvenne et al. 2010). MS exploits diffusion dynamics over an underlying graph structure to reveal multi-scale community organisation and their stability across time scales (see methods: Dynamical community detection).

The quasi-hierarchical community structure of the wards

Continuing with the lumped state network $\hat{\mathcal{M}}_2$, we apply MS and in Fig. 6 we show an apparent hierarchy of state node assignments to community partitions across Markov time t. We selected three points across Markov time (t_1,t_2,t_3) that exhibited robust community partitions (Additional file 1: Fig. S5). At longer time scales MS reveals coarser community partitions which show significant correspondence to hospital sites (Fig. 6). Specifically, at t_3 each cluster in the 3-way partition strongly corresponds to one of the three hospital sites. If we extend to even longer t we identify a 2-way partition where two hospitals are grouped almost exclusively into a single community (Additional file 1: Fig. S7). Notably, the hospital with wards grouped separately is the Tertiary site within the hospital trust which consists of speciality wards and appears to share fewer patients with the other two hospitals.

Moving towards shorter t within the MS analysis, which are expected to identify more granular structures of patient flow, we identity sub-structures largely contained *within* hospital sites, which overlap to a lesser extent between hospital sites (Fig. 6a). In some cases, these confer to buildings (we find 10 buildings that are over-represented in clusters at t_1), in other cases these confer to specialities (we find 7 specialities over-represented in clusters at t_1). Focusing initially on speciality, we find three specialities (Haematology, Cardiology, and Renal) that are over-represented within separate communities suggesting they are have a high degree of within speciality patient movement (Additional file 1: Fig. S8). However, as we increase t to reveal coarser partitions we see the more granular communities combine, bringing together previously distinct specialities such as Haematology or Renal into coarser partitions with other specialities, highlighting the zooming affect of MS as we change the t at which communities are observed. However, it is clear that the community structure is not entirely defined by specialities and the physical constraints imposed by buildings,

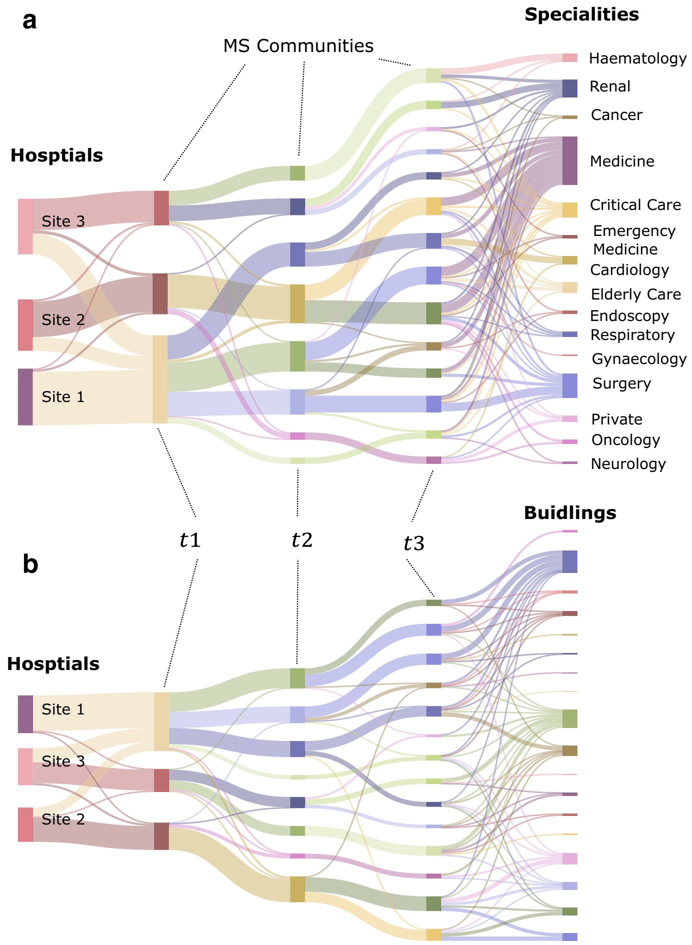

Fig. 6 Hierarchical breakdown of Markov Stability communities for three chosen scales t_1, t_2 and t_3 (optimal partitions chosen for their robustness after a detailed Markov Stability analysis, see Additional file 1: Fig. S4) and the relations of: (left) coarse partitions to Hospital sites; and (right) granular partitions to **a** specialities and **b** buildings

hospitals, and common movement patterns play a significant role and result in our observed communities (Fig. 6b).

Given that the majority of patients will move between specialities at some point during their journey through the hospital, it is expected that communities would not correspond exactly to ward specialities. Often this is attributable to patients seeking treatment for comorbidities, additional to their primary condition. The effect of such movements is a *memory effect* whereby patients will transition back to wards treating their main condition. In fact, several specialities primarily service these secondary conditions. An example is Medicine, a general class of ward that, as well as taking admissions also offers *general* treatment and support. Critical Care is another example with high expected memory effects, since it services patients from any ward if they deteriorate fast enough. Notably, we find that wards both in the Medicine and Critical Care specialities can be found within 10 different communities at t_1. Additionally, Surgery, another department that services multiple other wards, can be found in 9 different communities.

Overlapping community assignments

Although community detection generally is generally used to find 'disjoint' communities, *multiple community membership* is a well observed phenomenon, whereby a node may have multiple functions that it shares with different groups of nodes (Xie et al. 2013). Understanding that we are essentially clustering wards based on the movement patterns of patients, it is likely that different cohorts of CPE patients (e.g. with different comorbidities) have overlapping pathways. For instance, different cohorts of patients still require a set of common services and hence visit an overlapping set of wards (e.g. for admission, surgery, critical care, or renal dialysis). This phenomenon is well captured by memory networks, standard methods of community detection applied across the state network are able to reveal overlapping communities of nodes on the physical network. Additionally, the notation of granularity introduced by MS adds an interesting dimension to this problem, whereby the degree to which wards overlap communities can depend on the point Markov time. We can thus identify hospital wards which persistently overlap multiple communities across both granular and coarse time scales. These wards are of particular interest when developing Infection Prevention and Control strategies as they can play the role of network *bridges* and potential transmission hotspots.

At the most granular time scales, we find 48 wards with multiple community assignments (Additional file 1: Tables S2–S4). With increasing Markov time the total number of overlapping wards decreases; however, there exist several wards which are persistently overlap communities. We find 4 Renal wards and a single Elderly Care ward which have membership within each community of the 2-way coarse partition. Despite disappearing in the very coarser 2-way partition after $t > 12$, Critical Care, Medicine, and Surgery, as well as a single Elderly care ward also overlapped between communities. Since the most coarse partitions strongly corresponded to non-specialist hospital, and specialist hospital sites, it is likely that Critical Care and the Elderly care wards still play a strong connective role within connecting the two non-specialist hospital Sites 1&2.

Identifying the most central wards

In the previous section we identified nodes that were assigned to multiple communities, highlighting their critical role in the pathways of multiple cohorts of patients with differing patterns. We also used PageRank to identify the importance of wards in $\mathcal{M}_1$ and $\mathcal{M}_2$.

To allow for a more complete examination of ward importance, and to provide further investigation of $\hat{\mathcal{M}}_2$, we use Multiscale Centrality (MSC). MSC is a measure of centrality that enables us to identify nodes that are important in the network at different scales (Arnaudon et al. 2020). Following the same approach to compute centrality of the physical nodes, we compute MSC for each state node and then compute the sum of state node centralities across each physical node to generate a value of MSC for each ward.

Figure 7 shows the results of MSC computed for $\hat{\mathcal{M}}_2$. We find several wards that are central at all scales, implying that they are both highly connected locally (short time scales), and also important as global connectors/bridges (long time scales). Both Medicine 13 and 14 appear as central at all time-scales; Medicine 13 and 14 are both admission and readmission points into the hospital, where patients will be first identified as positive for CPE, and where they will return if readmitted. Additionally, we find four renal wards are central at all scales; note that three of these renal wards (Renal 1, Renal

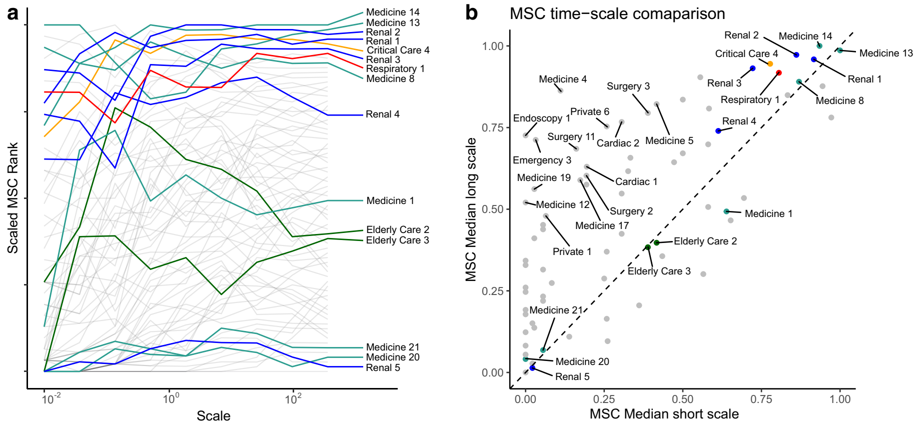

Fig. 7 Ranking of nodes by their multiscale centralities (MSC) in $\hat{\mathcal{M}}_2$. **a** Multiscale centrality ward rankings across all scales; all wards are plotted (light grey lines) with highlighted wards in colour. A high value of the scaled MSC Rank means that the wards are important. **b** A comparison of the MSC ranking at short scales versus long scales. The annotated wards are those with large differences in their PageRank in $\mathcal{M}_1$ versus $\mathcal{M}_2$. Note that at short scales, some nodes will have not been assigned a Multiscale centrality value and so share the bottom rank

2, and Renal 3) were also ranked highly in our previous analysis of M1 and M2 with Pagerank. Conversely, we also find wards that vary considerably in their importance across time-scales (Fig. 7b).

Conclusions

Analysis of patient movement can give valuable insights to understand disease dynamics and inform Infection Prevention and Control. Here we examined the common assumption of memoryless-ness in the movement of patients. To this end, we employed network analysis and compared networks with and without memory.

Models with memory have a substantially larger parameter space. We therefore constructed a simplified memory model based on a hybrid 'lumped' memory network, which retains the effect of memory in the patient trajectories but removes redundant or duplicate information. Such hybrid models are particularly useful for networks which exhibit different levels of memory depending on the localities. In this context, we extended previous work on lumping in memory networks in two ways: firstly, we defined a node feature vector that allows state nodes to be compared and lumped into meta nodes based on longer random walks; secondly, we proposed that lumping could be optimised by using prior knowledge with known communities which partially constrain patient pathways. This framework for constructing lumped memory networks is generalisable to both other hospital sites and other types of pathway data, where the underlying characteristics of the system predetermine many routes of movement. In the hospital context, patient trajectories are constrained by common pathways of patient treatment and care.

The lumped memory network formed the basis of subsequent investigation to detect communities of movement within our healthcare network. We used prior knowledge, including the hospital structure and specialities, to optimise the rates of lumping based on the Markov Stability graph clustering framework. As a result, we could highlight clusters of patient movement with higher-order memory and identify wards that appeared in multiple communities. The communities of patient movement divided the hospital sites quasi-hierarchically into sub-communities of wards that share patient flow. We found correspondence between community structures and known structures, such as hospital buildings or specialities; yet the communities also result from common pathways specific to certain cohorts of patients amongst this hospital group.

The higher-order network framework was specifically applied to analyse a large data set of CPE patient pathways collected in three large London hospitals. The analysis showed that the movement of hospitalised patients colonised with CPE displays substantial memory, i.e., ward transitions depend on previously visited wards. The presence of memory was identified by comparing node rankings with differing degrees of memory, as well as the statistics of a diffusion process on the resulting network models. Notably, we found that including memory in the network model increased the centrality of wards that are known clinically to be commonly visited by CPE patients (e.g. Renal) and decreased the centrality of wards that are less clinically visited amongst CPE patients (e.g. Paediatric). We note that the effect of memory can also be studied by analysing differences in the distribution of standard node centrality statistics (Additional file 1: Fig. S11). Memory also greatly affected local reachability. For example, the memory-less first-order model wrongly implied that patients could reach almost any ward within three ward moves after first entering the hospital. Our analysis, on the other hand, shows that accounting for where patients had previously been, dramatically restricts the possible set of subsequently visited wards. Hence, ignoring pathway memory in hospital patients can affect the outputs of commonly used network analytics tools, and potentially misrepresent the importance of hospital wards.

Understanding the constraints of patient movement can aid IPC. We showed that the ranking of wards and the likelihood of infected patients visiting particular wards was more accurate in the memory network than the memory-less network. In particular, we found that by extending beyond 1st-order memory, wards known to be associated with CPE were ranked more highly. Pinpointing at-risk wards is critical to focus IPC efforts and prevent transmission. Identifying communities of patients with distinct movement patterns, moreover, is valuable to cohort outbreaks within these communities and to prevent spread to other communities. Indeed, we found that the overlaps between communities revealed wards visited by almost all CPE patients (Renal wards) and wards visited commonly by the general patient population (Medicine, Surgery, and Critical Care wards). Such wards are prime targets of enhanced prevention efforts to reduce transmission.

Overall, our study highlights the role of memory in patient pathways. Most current analyses of patient pathways assume memoryless-ness as a first approximation. Here, we showed that ignoring memory may misidentify potential hubs of disease transmission. Our study gave some indication for memory beyond the second-order, however, we were limited in its detection due to the need of increasingly larger datasets. Future work incorporating higher-order effects may therefore give further insights into the precise nature of memory in patient movement. Our analysis suggests that informing policy based on traditional memory-less networks can miss key characteristics of movement patterns. For IPC this can mean missing transmission hubs and wrongly directing screening efforts to less critical locations, resulting in poor use of resources and lower efficacy. Our lumped memory network thus provides a framework for future patient-pathway analyses aimed at improving containment of CPE, and may be generalised to inform infection prevention and control of other HAIs.

Methods

Higher-order PageRank

PageRank is a measure of node importance or centrality within a network based on the incoming edges (Page et al. 1999). To obtain *Higher-order* PageRank we follow the derivation presented by Rosvall et al. (2014). PageRank is essentially computing the visitation probabilities to nodes over a network, determined by connectivity and weighting of those connections. In the context of a memory network, one can simply derive PageRank over the underlying state network for a model of arbitrary order k, and project the visitation probabilities back onto the physical nodes.

Firstly, we define the probability of finding a random walker on a given state node s at time $t + 1$ as

$$P(s_j; t + 1) = \sum_{s_i} P(s_i; t) p(s_i \to s_j), \tag{5}$$

where as before a state confers to a pathway of length k and transition probabilities are encoded by the transition matrix P.

Now, for any order k the higher-order generlisation of PageRank is simply the stationary solution to equation 5:

$$\pi(s_j) = \sum_{i} \pi(s_i) p(s_i \to s_j). \tag{6}$$

With $\pi(s_j)$ it is then trivial to return the physical node PageRank by summing over a physical nodes states:

$$\pi(k) = \sum_{j} \pi(s_j) = \sum_{k} \pi(s_j). \tag{7}$$

State lumping on local connectivity

Given a large set of trajectories, the problem arises that state node networks $\mathcal{M}_k$ can become very large and often contain redundancies. Not all pathways exhibit full transitive dependence, so it can often be desirable to reduce the model complexity by lumping together redundant state nodes. Redundancy of state nodes can lead to over-fitting when a physical node contains a number of similar states. Hence, we focus on lumping states nodes within the same physical node, forming so called 'meta state nodes' which also benefit from preserving the physical network structure (Lambiotte et al. 2019). For each lump, we reassemble all connections between two states nodes such that transition probabilities and connectivity are preserved (Edler and Bohlin 2017). In effect, 'lumping' state nodes together reduces the model complexity whilst retaining the transitive dependence of the original pathways.

In our approach, we lump together state nodes based on the similarity of visitation probabilities of the physical nodes. To do this we use the $S \times S$ state transition matrix P over k-steps and then sum the probabilities over the state nodes that compose each physical node. In the construction of P we add weighted self loops equivalent to a nodes total outflow weight $w_{s_i s_i} = \sum_{s_i} w_{s_i s_j}$ to derive P' with Eq. 3. This self loop conserves local flow across P', emphasising local connectivity when we subsequently determine distances across X.

We define the state node to physical node transition matrix X as the visitation probabilities of each state node to each physical node over k-steps, $X = P^k D$, where P is the state node transition matrix and D is the $S \times N$ state node to physical node indicator matrix. Each entry x_{ij} corresponds to the probability of transitioning from state node i to physical node j and thus provides a mapping from the higher order state node network to the physical node network. Here, we set $k = 3$ to incorporate a slightly larger range of local connectivity than previous works that use $k = 1$ (Edler and Bohlin 2017; Persson et al. 2016).

State nodes with similar local connectivity will exhibit similar probability distributions on the physical node network, therefore we can compute a similarity matrix between state nodes by computing the Wasserstein distance (Villani 2008) between vector rows of X which measures the distance for moving from one probability distribution to another. The similarity matrix was subsequently clustered using an agglomerative clustering method for lumping state nodes within physical node (Hastie et al. 2009).

In order to control the degree to which state nodes are lumped we employed a clustering rate r, which sets the number of final lumped state nodes that should be constructed for each physical node after completion of the lumping procedure. For example, lets consider a scenario where we have two physical nodes, one of which is composed of 10 state nodes and the second is composed of 20 state nodes. If we set the lumping rate $r = 0.2$, then after lumping the first physical node would have 2 lumped nodes after the procedure whereas the second physical node would have 4 lumped nodes. Increasing the lumping rate to $r = 0.8$ would mean physical nodes retain more of their states after the

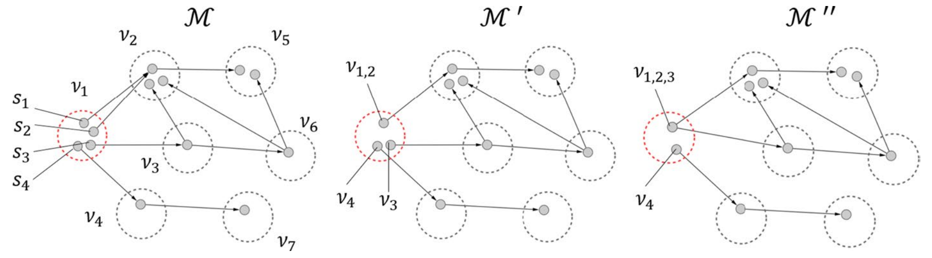

Fig. 8 State Lumping Example for a single physical node (v_1). Two possible lumpings $\mathcal{M}'$ and $\mathcal{M}''$ are visualised here over the state nodes (grey nodes) with the physical node mapped over (dotted circles surrounding state nodes). Here each lump merges the two most similar state nodes based on feature vectors capturing local visitation probabilities of $k = 2$ network steps

lumping, and for our example would result in those physical nodes having 8, and 16 final state nodes respectively.

Consider a simple illustrative lumping example in Fig. 8 which demonstrates the lumping process for a single physical node (circle of red dashed lines, v_1) and its constituent state nodes (grey circles within the red dashed circle) for different values of k. For the case $k = 1$ (see $\mathcal{M}'$ in middle of Fig. 8) only the nearest neighbours of each state node are considered and as such s_1 and s_2 will be lumped together first. The next lumping of state nodes is unclear given that both s_3 and s_4 have 1-step neighbors states in different physical node. However, as we increase k we explore more of the local network and at $k = 2$, in this example, it becomes clear that s_3 is more similar to s_1 and s_2. Hence for the second lumping, s_3 is merged with lumped meta node $s_{1,2}$ instead of s_4 (see $\mathcal{M}''$ in middle of Fig. 8).

Dynamical community detection

Dynamic community detection with Markovian assumptions can still be used to reveal structure in a memory network, simply by applying the same community detection algorithms to the higher-order network structure. $\mathcal{M}_k$, for $k > 1$, acts to constrain a walkers movement over the physical nodes within its state network connectivity. Hence, if we look for regions across $\mathcal{M}_k$ that conserve flow from a dynamical perspective, projecting the resultant communities back onto the physical nodes reveals overlapping communities constrained by the transitivity of the state network.

One such example for such a dynamical approach to community detection is Markov stability (MS) (Delvenne et al. 2010), which is the focus for this study. MS exploits diffusion dynamics over an underlying graph structure to reveal a multi-scale community organisation and has been show to be effective in a variety of applications in which multiple scales are expected to exist such as protein sub-structures (Peach et al. 2019a) or social behaviours (Peach et al. 2019b). Given a partition $\mathcal{P}$ of nodes into C non-overlapping communities with a $N \times C$ community indicator matrix $H_{\mathcal{P}}$ the time-dependent clustered autocovariance matrix in MS is given by,

$$R(t, H_{\mathcal{P}}) = H_{\mathcal{P}}^T \left[\Pi \, e^{t(M-I)} - \pi \pi^T \right] H_{\mathcal{P}}, \tag{8}$$

where the elements of the matrix $[R(t, H_{\mathcal{P}})]$ correspond to the probability of a random walker starting at node i and ending up in community c at Markov time t minus the probability of that happening by chance.

For an optimal partition $\mathcal{P}$, in which flow is trapped more than one would expect by random over t, we would expect a comparatively large Markov stability With the Markov stability as

$$r(t, H_{\mathcal{P}}) = \operatorname{trace} R(t, H_{\mathcal{P}}). \tag{9}$$

We aim to maximise $r(t, H_{\mathcal{P}})$ over the space of possible partitions $\mathcal{P}$ at a given Markov time t,

$$\mathcal{P}_{\max(t)} = \underset{\mathcal{P}}{\operatorname{argmax}} \, r(t, H_{\mathcal{P}}). \tag{10}$$

Whilst the optimisation of Eq. 10 is NP-hard, in practice, heuristics algorithms have been developed which are computationally efficient. Here we use the Louvain algorithm which has has been demonstrated to offer robust solutions at reasonable cost (Blondel et al. 2008).

Identifying partitions of interest over Markov-time

Given a set of partitions that are optimal at each Markov time we must still define which scales are representative or robust in respect to our system. In order to identify partitions of interest over time we look towards two robustness measures. Firstly, we look at consistency of partitions for single points in time, and secondly, we look for stable partitions across time.

To assess this consistency between $\mathcal{P}$ at Markov time t we can compute an information-theoretical distance between two alternate partitions $\mathcal{P}$ and $\mathcal{P}'$ is employed:

$$VI(\mathcal{P}_i(t), \mathcal{P}_j(t)) = \frac{2\Omega(\mathcal{P}, \mathcal{P}') - \Omega(\mathcal{P}) - \Omega(\mathcal{P}')}{\log(n)}, \tag{11}$$

where $\Omega(\mathcal{P})$ is the Shannon entropy, $\mathcal{P}_\mathcal{C}$ being the relative frequency of finding a node in community $\mathcal{C}$ in partition $\mathcal{P}$.

Then to quantify consistency at Markov time t we compute the average variation of information of all solutions:

$$\langle V(t) \rangle = \frac{1}{l-1} \sum_{i \neq j} VI(\mathcal{P}_i(t), \mathcal{P}_j(t)). \tag{12}$$

For the case that optimisations return near identical partitions $\langle V(t) \rangle$ will be small, which indicates robustness of the partition at t. Hence over t we search for partitions with low $\langle V(t) \rangle$.

Relevant partitions should also be remain consist across regions of Markov time. Such persistence is indicated both by a plateau in the number of communities over t and a low value or plateau of the cross-time variation of information:

$$VI(t, t') = VI(\widehat{\mathcal{P}}(t), \widehat{\mathcal{P}}(t')). \tag{13}$$

Multi-scale centralities

For identification of central nodes we use Multiscale Centrality, that enables us to identify nodes that are central at different scales within the network (Arnaudon et al. 2020). Multiscale centrality leverages the presence of 'overshooting' peaks that appear in diffusion processes on the graphs. For a more detailed description of overshooting peaks, see (Peach et al. 2020). Central nodes are defined as a node, i that breaks the triangle inequality for a pair of nodes j, k,

$$\Delta_{ij,\tau} := t^*_{ij,\tau} + t^*_{ik,\tau} - t^*_{jk,\tau} \leq 0,$$

where $t_{ij,\tau}$ is the Markov time at which an overshooting peak appears at node j given the diffusive process of an initial delta function at node i which is allowed to diffuse up to Markov time τ.

The diffusion process underlying Multiscale centrality acts as a scaling factor that allows us to identify nodes that are central at different scales of the network structure. For example, some nodes may be locally central (with high degree) or might be globally central (high closeness). Thus we produce a ranking of nodes as a function of Markov time τ of the diffusion process. For further details on this methodology, see Arnaudon et al. (2020).

For each state node we can compute the Multiscale centrality. In an identical manner to Higher-order PageRank (see "Higher-order PageRank" section), we can then compute a physical node centrality by summing the multiscale centrality over the constituent state nodes.

Abbreviations

AMR: Antimicrobial resistance; HAI: Healthcare-associated infection; IPC: Infection prevention and control; CPE: Carbapenemase-producing Enterobacteriaceae; MS: Markov stability; MSC: Multiscale centrality.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s41109-021-00376-5.

Additional file 1: Fig. S1. State networks of $\mathcal{M}_1$ and $\mathcal{M}_2$. **Fig. S2**. PageRank difference between $\mathcal{M}_1$ and $\mathcal{M}_2$ over specialities and buildings. **Fig. S3**. Optimisation of clustering rate for state lumping. **Fig. S4**. Markov Stability run statistics. **Fig. S5**. Markov stability community partitions. **Fig. S6**. Variation of Information between hospital structures in community partitions. **Fig. S7**. 2-Way community partition to hospital site. **Fig. S8**. Hospital wards overlapping communities across Markov stability partitions. **Figs. S9, S10**. Multiscale centrality model comparison. **Fig. S11**. Distribution of node statistics. **Table S1**. Cross validation ranking significance. **Tables S2–S4**. Examining the overlap between Markov stability community partitions and known hospital structures.

Acknowledgements

We wish to thank Eleonora Dyakova for help with accessing data. We would also like to thank Frankie Bolt, and Juliet Allibone for supporting the final publication process.

Authors' contributions

ACM and RLP performed the computational analyses. ACM, RLP, and MB contributed to methodological development. SM, AH, and FD provided interpretation of data and results. ACM, RLP, AYW, and MB wrote the paper. AYW, AH, and MB provided project supervision. All authors read and approved the final manuscript.

Funding

ACM was supported in part by a scholarship from the Medical Research Foundation National PhD Training Programme in Antimicrobial Resistance Research (MRF-145-0004-TPG-AVISO), as well as by the National Institute for Health Research Academy. FD is supported by the Medical research council Clinical academic research partnership scheme. AH is a National Institute for Health Research Senior Investigator. AH is also partly funded by the National Institute for Health Research Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Infections in partnership with Public Health England, in collaboration with, Imperial Healthcare Partners, University of Cambridge and University of Warwick. ACM, RP, and MB acknowledge funding from EPSRC Grant EP/N014529/1 to MB, supporting the EPSRC Centre for Mathematics of Precision Healthcare. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health and Social Care or Public Health England.

Availability of data and materials

The datasets generated and analysed during the current study are not publicly available to protect anonymity of included hospital patients. The repository for Multiscale centrality can be found at https://github.com/barahona-research-group/MultiscaleCentrality.

Declarations

Ethics approval and consent to participate

Patient pathway data was collected from the central business intelligence system and fully pseudanonymised before analysis, in accordance with ethics 15_LO_0746.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Department of Mathematics, Imperial College London, London, UK. [2]Department of Infectious Disease, Imperial College London, London, UK. [3]Present Address: School of Informatics, University of Edinburgh, Edinburgh, UK. [4]Imperial College Healthcare NHS Trust, London, UK. [5]Department of Neurology, University Hospital Würzburg, Würzburg, Germany.

Received: 22 September 2020 Accepted: 20 April 2021
Published online: 03 May 2021

References

Arlot S, Celisse A et al (2010) A survey of cross-validation procedures for model selection. Stat Surv 4:40–79. https://doi.org/10.1214/09-SS054

Arnaudon A, Peach RL, Barahona M (2020) Scale-dependent measure of network centrality from diffusion dynamics. Phys Rev Res 2:033104. https://doi.org/10.1103/PhysRevResearch.2.033104

Balcan D, Vespignani A (2011) Phase transitions in contagion processes mediated by recurrent mobility patterns. Nat Phys 7(7):581–586. https://doi.org/10.1038/nphys1944

Bean DM, Stringer C, Beeknoo N, Teo J, Dobson RJB (2017) Network analysis of patient flow in two UK acute care hospitals identifies key sub-networks for A&E performance. PLoS ONE 12(10):1–16. https://doi.org/10.1371/journal.pone.0185912

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Bonomo RA, Burd EM, Conly J, Limbago BM, Poirel L, Segre JA, Westblade LF (2018) Carbapenemase-producing organisms: a global scourge. Clin Infect Dis 66(8):1290–1297. https://doi.org/10.1093/cid/cix893

Bühlmann P, Wyner AJ et al (1999) Variable length Markov chains. Ann Stat 27(2):480–513. https://doi.org/10.1214/aos/1018031204

Cawley GC, Talbot NLC (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 11:2079–2107

Chierichetti F, Kumar R, Raghavan P, Sarlos T (2012) Are web users really Markovian? In: Proceedings of the 21st international conference on World Wide Web. WWW'12, pp 609–618. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2187836.2187919

Delvenne J-C, Yaliraki SN, Barahona M (2010) Stability of graph communities across time scales 107(29):12755–12760. https://doi.org/10.1073/pnas.0903215107

Donker T, Wallinga J, Slack R, Grundmann H (2012) Hospital networks and the dispersal of hospital-acquired pathogens by patient transfer. PLoS ONE 7(4):1–8. https://doi.org/10.1371/journal.pone.0035002

Edler D, Bohlin L et al (2017) Mapping higher-order network flows in memory and multilayer networks with infomap. Algorithms 10(4):112. https://doi.org/10.3390/a10040112

Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782. https://doi.org/10.1214/09-SS0540

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin

Heath MF, Vernon MC, Webb CR (2008) Construction of networks with intrinsic temporal structure from UK cattle movement data. BMC Vet Res 4(1):11. https://doi.org/10.1214/09-SS0541

Interagency Coordination Group on Antimicrobial Resistance (2019) No time to wait: securing the future from drug-resistant infections. Technical report. https://doi.org/10.1214/09-SS0542. Accessed 30 Aug 2020

Jääskinen V, Xiong J, Corander J, Koski T (1983) Sparse Markov chains for sequence data. Scand J Stat 41(3):639–655. https://doi.org/10.1214/09-SS0543

Kareiva P, Shigesada N (1983) Analyzing insect movement as a correlated random walk. Oecologia 56(2–3):234–238. https://doi.org/10.1214/09-SS0544

Kendall MG (1938) A new measure of rank correlation. Biometrika 30(1–2):81–93. https://doi.org/10.1214/09-SS0545

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/10.1214/09-SS0546

Logan LK, Weinstein RA (2017) The epidemiology of carbapenem-resistant enterobacteriaceae: the impact and evolution of a global menace. J Infect Dis 215(suppl-1):28–36. https://doi.org/10.1214/09-SS0547

Masuda N, Porter MA, Lambiotte R (2017) Random walks and diffusion on networks. Phys Rep 716–717:1–58. https://doi.org/10.1214/09-SS0548

May RM, Lloyd AL (2001) Infection dynamics on scale-free networks. Phys Rev E 64:066112. https://doi.org/10.1214/09-SS0549

Mucha PJ, Richardson T, Macon K, Porter MA, Onnela J-P (2010) Community structure in time-dependent, multiscale, and multiplex networks 328(5980), 876–878. https://doi.org/10.1126/science.1184819

O'Neill J (2016) Tackling drug-resistant infections globally: final report and recommendations. Technical report, Review on Antimicrobial Resistance. https://doi.org/10.1103/PhysRevResearch.2.0331040. Accessed 01 Sept 2020

Organization WH et al. (2002) Prevention of hospital-acquired infections: a practical guide. Technical report, Geneva, Switzerland: World Health Organization. https://doi.org/10.1103/PhysRevResearch.2.0331041. Accessed 09 Sept 2020

Otter JA, Mookerjee S, Davies F, Bolt F, Dyakova E, Shersing Y, Boonyasiri A, Weiße AY, Gilchrist M, Galletly TJ, Brannigan ET, Holmes AH (2020) Detecting carbapenemase-producing Enterobacterales (CPE): an evaluation of an enhanced CPE infection control and screening programme in acute care. J Antimicrob Chemother 75(9):2670–2676. https://doi.org/10.1103/PhysRevResearch.2.0331042

Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web. Technical report 1999-66, Stanford InfoLab. Previous number=SIDL-WP-1999-0120. https://doi.org/10.1103/PhysRevResearch.2.0331043
Palla G, Páll N, Horváth A, Molnár K, Tóth B, Kováts T, Surján G, Vicsek T, Pollner P (2017) Complex clinical pathways of an autoimmune disease. J Complex Netw 6(2):206–214. https://doi.org/10.1103/PhysRevResearch.2.0331044
Pastor-Satorras R, Vespignani A (2001a) Epidemic spreading in scale-free networks. Phys Rev Lett 86:3200–3203. https://doi.org/10.1103/PhysRevResearch.2.0331045
Pastor-Satorras R, Vespignani A (2001b) Epidemic dynamics and endemic states in complex networks. Phys Rev E 63:066117. https://doi.org/10.1103/PhysRevResearch.2.0331046
Peach RL, Saman D, Yaliraki SN, Klug DR, Ying L, Willison KR, Barahona M (2019a) Unsupervised graph-based learning predicts mutations that alter protein dynamics. https://doi.org/10.1101/847426
Peach RL, Yaliraki SN, Lefevre D, Barahona M (2019b) Data-driven unsupervised clustering of online learner behaviour. NPJ Sci Learn 4(1):1–11. https://doi.org/10.1103/PhysRevResearch.2.0331047
Peach RL, Arnaudon A, Barahona M (2020) Semi-supervised classification on graphs using explicit diffusion dynamics. Found Data Sci 2(1):19. https://doi.org/10.1103/PhysRevResearch.2.0331048
Persson C, Bohlin L, Edler D, Rosvall M (2016) Maps of sparse Markov chains efficiently reveal community structure in network flows with memory. arXiv preprint https://doi.org/10.1103/PhysRevResearch.2.0331049
Poletto C, Tizzoni M, Colizza V (2013) Human mobility and time spent at destination: impact on spatial epidemic spreading. J Theor Biol 338:41–58. https://doi.org/10.1038/nphys19440
Prestinaci F, Pezzotti P, Pantosti A (2015) Antimicrobial resistance: a global multifaceted phenomenon. Pathog Glob Health 109(7):309–318. https://doi.org/10.1038/nphys19441
Rissanen J (1983) A universal data compression system. IEEE Trans Inf Theory 29(5):656–664. https://doi.org/10.1038/nphys19442
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(1):1–13. https://doi.org/10.1038/nphys19443
Salnikov V, Schaub MT, Lambiotte R (2016) Using higher-order Markov models to reveal flow-based communities in networks. Sci Rep 6:23194. https://doi.org/10.1038/nphys19444
Schaub MT, Lambiotte R, Barahona M (2012) Encoding dynamics for multiscale community detection: Markov time sweeping for the map equation. Phys Rev E 86(2):026112
Scholtes I (2017) When is a network a network? multi-order graphical model selection in pathways and temporal networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1037–1046. https://doi.org/10.1145/3097983.3098145
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423
Singer P, Helic D, Taraghi B, Strohmaier M (2014) Detecting memory and structure in human navigation patterns using Markov chain models of varying order. PLoS ONE 9(7):1–21. https://doi.org/10.1038/nphys19445
Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility 327(5968):1018–1021. https://doi.org/10.1126/science.1177170
Struelens MJ (1998) The epidemiology of antimicrobial resistance in hospital acquired infections: problems and possible solutions. BMJ 317(7159):652–654. https://doi.org/10.1038/nphys19446
Villani C (2008) Optimal transport: old and new, vol 338. Springer, Berlin
Xie J, Kelley S, Szymanski BK (2013) Overlapping community detection in networks: the state-of-the-art and comparative study. ACM Comput Surv. https://doi.org/10.1038/nphys19447

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.