Imperial College of Science, Technology and Medicine

Department of Mathematics

Second-Order Time-Unfolded Models of Contagion Applied to Antimicrobial Resistance and the COVID-19 Pandemic

Matthew Stuart Packham

Submitted in part fulfilment of the requirements for the degree of

MSc Applied Mathematics the Diploma of Imperial College, September 2020/21

Abstract

The dynamics of disease transmission is reliant upon the time distribution of agent contacts. Attributes such as recovery time, infectious periods or incubation times restrict the ability of a disease to pass between agents. In this thesis the notion of a second-order time-unfolded graphical model known as an Event Graph which allows connectivity determined by time-respecting paths, where the time between contacts has an upper limit, is studied in relation to contagion.

We apply this representation to two separate pathogens which stem from the Imperial College London NHS Trust, Carbapenemase Producing Enterobacteriace (CPE) an antibiotic resistant bacteria which is causing a critical public healthcare problem globally and SARS-Cov-2, the virus responsible for the current coronavirus disease (COVID-19) pandemic.

Our model explicitly captures the patterns of interactions among patients within the Imperial College London NHS Trust and analysis of the contact sequences (weakly-connected components of the Event Graph) leads us to cluster contact structures, based on the similarities of temporal and topological features. These contact clusters are then analysed to provide insight on how specific diseases may be facilitated by their structure. We conclude with recommendations to the Trust of ways to combat future outbreaks of these diseases.

Declaration

The work contained in this thesis is my own work unless otherwise stated.

Acknowledgements

I would like to express my profound gratitude and appreciation to:

- My supervisor Dr Mauricio Barahona, Chair in Biomathematics at Imperial College, for allowing me the opportunity to undertake this thesis and his continued support.

- My second supervisor Dr Robert Peach, Honorary Research Fellow at the Department of Brain Sciences Imperial College, for his valuable guidance and expertise.

- My third supervisor Mr Ashleigh Myall, Research Postgraduate at the Centre of Mathematics for Precision Healthcare Imperial College, for whom I am indebted for his adept, sincere and valuable insights and encouragement throughout.

- My family and friends for their endless support and love.

- The Imperial College London NHS Trust for their generosity in approving the study ethics for the data used in this thesis.

‘It will happen but it will take time.’

John Bowlby

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Mathematics has a rich and well-established place in epidemiology [2] [17]. It is an invaluable epidemiological tool as it allows us to gain an insight into how specific strategies, variables and conditions can be used effectively to control the spread of an infectious disease, without the controversy of conducting unethical or impractical real experiments. Contact networks have been used to effectively to capture the diverse interactions that underlie the spread of disease within a population [37]. For example, *Newmann et al* [38] found that the caregivers and the extent of their contact patterns were fundamental to the control and prevention of mycoplasma, a bacteria which causes pneumonia, outbreaks within hospitals in the US. In a contact network, each person (or location) translates into a vertex, and contacts among people (or locations) translate into edges that connect appropriate vertices [43].

In this thesis, we build upon the intuition behind a contact network for capturing the spread of diseases by extending a temporal contact network to the notion of a second-order time-unfolded graphical model known as an Event Graph [34]. The aim is to gain invaluable insight into possible routes of transmission of two separate diseases, namely Carbapenemase Producing Enterobacteriaceae (CPE) (an antibiotic resistance bacteria) and SARS-CoV-2 (COVID-19), within the Imperial College London NHS Trust. We also offer our recommendations, based on our analysis of the contact sequences (components) and their composition, of the Event Graph, on how to tackle the spread of these diseases within the Trust.

1.1 Carbapenemase Producing Enterobacteriaceae

Microorganisms have evolved mechanisms to evade demise from antimicrobial, such as antibiotics and antivirals, through a Darwinian selection process. These include preventing entry of or exporting the antimicrobial, producing enzymes which destroy or modify the drug or mutations that change the antimicrobial target [20]. One such example, where antimicrobial resistance has had an effect on human health is the development of resistance to Carbapenems, an antibiotic which falls within the beta-lactam family. Carbapenems, due to their unique molecular structure (the presence of a carbapenem together with the beta-lactam ring), allow for a broad spectrum of antibacterial activity [32]. They are exceptionally stable against most beta-lactamases (enzymes that inactivate beta-lactams) including ampicillin and carbenicillin, which are used to treat infections such as meningitis (infection of the membranes that surround the brain and spinal cord) and infections of the throat, lungs and reproductive organs [32].

Carbapenemase-production in the family Enterobacteriaceae (CPE) are a specific gram-negative bacteria which are resistant to the Carbapenem class of antibiotics. CPE are particularly concerning due to their ease of exchanging genetic information, high nosocomial transmissibility and likelihood of simultaneously being resistant to other antibiotic classes [4]. As Carbapenems present fewer adverse effects than other antibiotics in the beta-lactam family and are highly effective against many bacterial species, Carbapenems are used for treating servere bacterial infectons often as a last resort [41]. Therefore, the emergence and rapid spread of Carbapenem resistance globally, mainly among Gram-negative bacteria, has resulted in a global public healthcare problem of major importance. In fact, CPE is listed as one of three pathogens in the critical category (highest priority) in the World Health Organisation's (WHO) priority pathogen list for Research and development of new antibiotics [51].

1.2 Coronavirus Disease

Coronaviruses (CoVs) are a family of viruses that cause respiratory and intestinal illnesses in humans and animals [7]. Initially it was thought that CoVs only infected animals until in 2002 the emergence of severe acute respiratory syndrome (SARS) outbreak in Guangdong, China which infected 8000 people and lead to 776 deaths. Then a decade later, another coronavirus, known as the Middle East respiratory syndrome coronavirus (MERS-CoV) caused an endemic in Middle Eastern countries. The World health organization (WHO) reported that MERS-CoV infected more than 2428 individuals and resulted in 838 deaths [49]. Since December 2019, the world has been battling another coronavirus, Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It is the virus responsible for the current coronavirus disease (COVID-19) pandemic, which was first identified in Wuhan, China, following reports of serious pneumonia [57] [60].

COVID-19 is a ruthless virus affecting all ages and genders with symptoms ranging from mild, self-limiting respiratory tract disease to aggressive pneumonia, multiple organ dysfunction, and in many cases death. [21]. As of the date of this thesis, there are over 225 million reported cases globally with over 4.6 million deaths [56]. It is therefore imperative that research is undertaken in an attempt to combat and better understand the transmission of this virus. Not only for the current epidemic but also to put us in a better standing for possible future novel CoVs.

1.3 Contributions

First we describe the necessary background theory including a literature review, to aid the reader. We also describe the datasets we were generously allowed rights to. In Chapter 3, we propose our percolation of our temporal contact network to that of an Event Graph as a way of a higher-order description. We describe the steps taken in the decomposition of our Event graph, including the method taken to optimise the Δ-t parameter essential for modelling the epidemiological dynamics. Upon creating our Event Graph, in Chapter 5 we analysis its temporal, topological and behavioural characteristic to bring about invaluable insights into disease transmission. Finally, in Chapter 6, we provide our recommendations to the Imperial College London Trust and suggest areas of future work.

Chapter 2

Methods and Theory

2.1 Networks

Before defining temporal networks and event graphs we first must start from the very beginning and define what is meant by a graph also known as a network.

Definition 2.1 ***Graph (Network).***
A Graph or Network is an ordered pair of disjoint sets (V, E) *such that* E *is a subset of the set* V *of unordered pairs of* V*. The set* V *is the set of vertices and* E *is the set of Edges. An edge* $\{i, j\}$ *is said to join the vertices* i *and* j *and denoted* ij*. If* $ij \in E$*, then* i *and* j *are said to be adjacent [3].*

Networks are very versatile and as such are used in a variety of fields [40]. There are four broad categories of networks Technological, Information, Social and Biological. Examples of Technological Networks are the British Telecommunications telephone network, airline routes and distribution networks such as oil-pipelines. Information Networks are slightly more abstract and consist of the structures of bodies of information, these include the World Wide Web or the network of citations linked between different academic studies. Social networks are networks where the Nodes represent people and the edges between them are social interactions (friendship, communication, interactions). The classic examples are Social Network platforms such as Twitter or Facebook. The fourth and final category is Biological Networks, which as the name suggest occur in a biological setting. Examples include the network of Neurons in the brain,the circulatory system and metabolic networks.

In many cases we use networks in our day-to-day activities without even knowing. For example, in Google Maps, various locations are represented as vertices and roads are represented as edges. Graph theory is used to find the shortest path between two nodes, representing the quickest route to your destination [10].

A graph as defined in Definition 2.1, has edges which do not change with time and remain present/on throughout the network, graphs with this property are known as static. The restriction of a network to being static is not always true [25]. Despite static networks providing a useful abstraction to complex systems, many networks are intrinsically dynamic and are in fact reliant upon their relationship with time [19]. Consider the reachability properties of airline travel out of Heathrow Airport. For passengers on in-direct flights the scheduling of possible paths to your destination is time-dependent and therefore we cannot realistically map this to a static network - which would suggest there is always a path to your destination no matter the time! Therefore, static networks can be misleading and suggest paths which do not exist in reality, because of time restrictions, when dealing with a network which is intrinsically dynamic. To deal with dynamical systems and capture their rich temporal dimension, we introduce the notion of a Temporal Network:

Definition 2.2 ***Temporal Network.***

We consider temporal networks as a sequence of temporal events E. Let $V \subset \mathbb{N}$ be a set of interacting nodes, and $T \subset \mathbb{R}_{\geq 0}$ a non-empty ordered set of interaction times, then the temporal network is defined as the tuple $G = (V, E, T)$ where $E \subset V^2 \times T$ consisting of a sequence of time ordered temporal events $(e_i)_{i=1}^{M}$. An individual event $e_i = (u_i, v_i, t_i) \in E$ corresponds to an interaction of node u_i with node v_j at time t_i (here assuming interaction is instantaneous and that each t_i is distinct). An intuitive way to visualise a temporal network is the aggregation of individual events [35].

Temporal networks allow us to capture information about the dynamics which otherwise would not exist in the static network [19]. Similar to the airline example proposed earlier, consider path-counting. A temporal network, where edges are labelled with the time at which they occur, restrict the possible paths present in the network, as a path must follow a time-ordered sequence of events [35]. Whereas, when considering only the aggregated (non-temporal) static network, where edge labels are removed - see Figure 2.1, then there could exists paths which would not otherwise exist in the underlying dynamical system. This has the potential risk of overestimating the number of possible paths through the network. Which, when considering how a disease spreads across a network of contacts, can mislead epidemiological investigations. This is one of the many reasons why temporal networks are far more realistic and provide a natural way of modelling networks with underlying epidemiological dynamical behaviour.

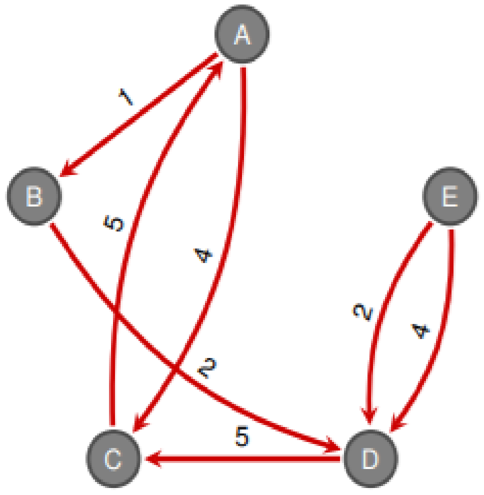


Figure 2.1: An example temporal network where edges are labelled with the time at which they appear (in this case, instantaneously). When considering the non-temporal network, remove edge labels, Figure taken from *Mellor* [35].

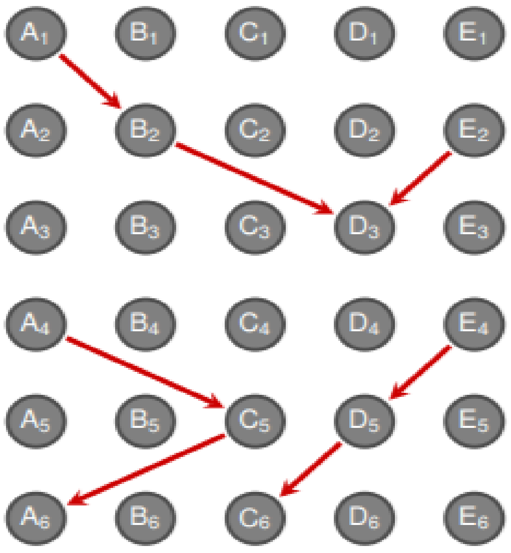


Figure 2.2: An example time-unfolded model where nodes in the temporal network are replace by time-indexed copies of themselves. Edges are then created between a time-index node and another time-indexed node that is strictly further ahead in time, Figure taken from *Mellor* [35]

One way to ensure we preserve the temporal ordering of paths is by considering the time-unfolded graphical representation. An intuitive static representation of temporal networks which in previous studies has shown to be effective [44] [48]. The key idea of this two-dimensional static representation is to arrange all nodes on a horizontal dimension, while unfolding time to an additional vertical dimension, as illustrated in Figure 2.2 [47]. Time-unfolded graphs are particularly useful as they are both directed and acyclic. Therefore are amenable to traditional static network methods.

Another common temporal network representation is the concept of a second-order model. In this representation the edges of the original static network become the nodes of the new model, and edges are connected to other edges if they have a node in common [35]. Second-order models are commonly referred to as memory networks because of their use in modelling random walks with memory [28]. The order can be extended to the k^{th} - dimension. Examples of k^{th} - order models include De Bruijin graphs [9] and Hashimoto graphs [16].

In this project we follow the definition of an Event Graph as outlined by *Mellor* [35], which combines the notion of a second-order model with that of a time-unfolded model.

2.2 Event Graph

We consider a Temporal Network defined by a sequence of temporal events $(e_i)_{i=1}^{M}$ where each event is a triplet of the form $e_i = (u_i, v_i, t_i)$. In addition, we make the assumption that a node may participate in only one event at a time i.e. that events are dyadic and consist only of pairwise interactions.

To define the Event Graph we first need to be able to relate two events in a meaningful way, which captures the relationship between events (vertices) and the temporal proximity of events. This will be represented by edges between events (vertices) and the way we relate edges will be by considering that of Δ-t Adjacency, which we define in Definition 2.4. For any two events we can examine the number of shared nodes (those interacting in an event) and the time between the two events occurring, defined as the Inter-Event time.

Definition 2.3 ***Inter-Event time.***

The Inter-Event time τ between two events $e_i = (ui, vi, ti)$ and $e_j = (uj, vj, tj)$ is given by:

$$\tau(e_i, e_j) = \begin{cases} t_j - t_i & t_j > t_i \\ 0 & otherwise \end{cases}$$

Definition 2.4 ***Δ-t Adjacency***

Two time-ordered events ei, ej are said to be Δ-t Adjacent if they share at least one node ($\{ui, vi\} \bigcap \{uj, vj\} = \emptyset$) and the time between the two events (The Inter-Event time) is no greater than t. $(0 < t_j - t_i \leq t)$

Adjacency makes intuitive sense in that information can only be transmitted between two events if there is one or more common nodes (those interacting in an event) where that information can persist.

Now we define the notion of an event graph.

Definition 2.5 ***(Δ-t) Event Graph***

For a temporal network $G = G(V, E, T)$, where $V \subset \mathbb{N}$ is a set of nodes, $E \subset V^2 \times T$ the set of temporal events and $T \subset \mathbb{R}_{\geq 0}$ a set of interaction times. Then the Δ-t Event Graph, is a directed graph $\zeta = \zeta(\nu, \xi)$ with $\nu = E$ and $\xi \subset \nu \times \nu$. The graph is defined such that there is a vertex for each event in E and each vertex is connected to the subsequent Δ-t Adjacent event of each node in that event.

More precisely, let

$$S(u_i) = \{k | (u_i \cap u_k, v_k = \emptyset) \cap (0 < t_k - t_i \leq t)\}$$

be the set of subsequent Δ-t adjacent events for the node u_i with the equivalent set defined for v_i.

The set of edges in the Δ-t Event Graph is then given by:

$$\xi = \{(e_i, e_j) | (j = \min\{S(u_i)\} \cup (j = \min\{S(v_i)\}\}$$

This construction results in a directed acyclic graph with a maximal in/out degree of at most two [34]. Weighting these edges with the time between the two events (inter-event time, $\tau(e_i, e_j)$) gives rise to a weighted, directed graph of temporal events.

Furthermore, two events are Δ-t-connected if they can be joined by a sequence of pairwise Δ-t-adjacent events [34]. That is, if two events in the Event Graph are connected via a (time restricted) sequence of Δ-t-adjacent events, then they are considered Δ-t-connected.

An Event Graph can be thought of as a dual network, where the events of the original network are now the vertices of the Event Graph and edges are connected to other edges if they have a node in common [18]. Therefore, if an event consists of Patient A interacting with Patient B (AB), then at a later time Patient B interacts, in another event, with Patient C (BC), then an edge would connect these two interactions, provided Δ-t adjacency is met, because they both have a node in common, that being Patient B.

2.3 Literature Review

The use of an Event Graph representation as a way of a higher-order description of a temporal network has been the focus of many studies. The concept of a second-order (dual) time-unfolded model known as an Event Graph was first introduced in 2017 by Dr. Andrew Mellor of the University of Oxford in his study *Mellor* [34]. Where he introduced the concept of this static, behavioural representation of a temporal network. Crucially, there exists the assumption of dyadic (pairwise) interactions between events. He provided new methods, through considering the distributions of the inter-event times and the two-event temporal motifs in union, to characterize the behaviour of individuals in temporal networks as well as providing a natural decomposition of the network.

The concept of the Event Graph was further formulized by *Kivela et al.* [26] in 2018. Where they studied the percolation on the (weighted) event graphs and in the underlying temporal networks, through simulated and real-world networks. Furthermore, they showed that this type of temporal-network percolation is analogous to directed percolation.

The most recent advancements in the area of Event Graphs comes again from Dr. Mellor in his paper *Mellor* [35] in 2019. Where he crucially introduces the first way to model non-dyadic interaction events at a network level, through what he defined as a Hyper-Event Graph. Hyper-Event Graphs allow nodes to interact with multiple other nodes, however the assumption that a node can only participate in one event at a time still holds.

2.4 Data Abstract

In this project we will consider two independent (temporal) data sets, one relating to CPE and the other to COVID-19. Both data sets stem from the Imperial College London NHS Trust where study ethics were granted and approved. All patient pathway data was extracted from routinely available electronic health data and fully pseudo-randomised before analysis in accordance with ethics 15_LO_0746. The data sets consist of patients for whom have tested positive for their respective condition. CPE was tested through a microbiological test, whereas COVID-19 was determined via a virology test. The data sets comprises patient interactions (Source and Target), the day in which the interaction occurred and the corresponding location within the Trust.

In Table 2.1 we have outlined the dataset statistics for both the CPE and Covid-19. We have included the total number of patients, the total number of events (interactions between patients), total duration of the data ([first event, last event]) and the total number of locations (locations are represented by wards across the trust). To reiterate, the data sets consist of only patients for whom have tested positive for their respective condition and we have omitted the general patient population.

	Patients	Total Events	Total Duration (Days)	Number of Locations
CPE	887	1608	659 [2020-06-10, 2018-08-21]	58
COVID-19	2159	11,146	266 [2020-11-01, 2021-02-09]	60

Table 2.1: Summary statistics for CPE and COVID-19 datasets

Chapter 3

Results: Event Graph Construction

In this section we describe the methods used to model the temporal interactions between patients who tested positive for CPE and COVID-19 respectively through the representation of an Event Graph. We describe the steps taken in the decomposition of our Event graph, including the method taken to optimise the value of the Δ-t parameter, essential for modelling the underlying epidemiological dynamics. Finally, we also set out the features that will be used to express the similarities between the contact sequences (components) of our Event graph and look to interpret these in relation to the underlying epidemiological structure.

3.1 Event Creation and Temporal Decomposition

Our aim is to model each of our temporal networks as an Event Graph. An Event Graph is a second-order time-unfolded model of the temporal network, where the events of the original temporal network are the nodes of the Event Graph [35]. Events are linked to the subsequent events for each node (patient) in that event, resulting in a directed acyclic graph with a maximal in/out degree of two. Weighting these edges with the time between the two events (inter-event time) gives rise to a weighted, directed graph of temporal events which we call the Event Graph.

The decomposition of the temporal network using an Event graph requires us to choose a value of the parameter Δ-t to threshold the time between adjacent events occurring for us to consider them connected. There is very little literature on the process of tuning the parameter Δ-t and very little has been done to investigate how the Event Graph changes with variations in Δ-t. In other studies, the authors have considered the sensitivity of the final results to variations in the parameter [34] [52]. In the following we specify our method for choosing the optimal value of the Δ-t parameter.

3.1.1 CPE Modelling Considerations

In the context of this thesis, Δ-t represents the maximum time in which a patient can interact with another patient with the possibility of pathogen transmission to occur over that event. Despite the lack of literature on tuning the Δ-t parameter, in the context of this thesis, there exists clinical research into the incubation time of CPE, which according to *Mo et al.* [39] is given by 86 days (within a 95% credible interval). Therefore, since CPE has a mean carriage duration of 86 days, this will be our value of Δ-t. As interactions between patients which occur after a duration of 86 days lack the potential to transmit CPE, we wish to ignore these and only look for contact events where disease transmission could occur.

Despite the clinical research, we would also like a way of tuning the Δ-t parameter ourselves based only on how descriptive features of the event graph change for variations in Δ-t. Therefore we introduce our own method to tune the Δ-t parameter. To our knowledge, this is the first use of this method. Since we are dealing with an unsupervised learning approach, we will examine how variations in Δ-t effect the temporal topological structure of the Δ-t Event Graph, by considering four separate temporal and topological measures on the event graph with the aim to find the optimal value of the Δ-t parameter. We consider the following four separate temporal and topological measures on the event graph:

- Average length duration
- Average size of components
- The number of connected component
- The number of edges

For comparison, we normalised and plotted how these changed with variations in Δ-t. Using the unsupervised heuristic known as the 'Elbow Method' we attempt to quantify the optimal value of Δ-t such that the fit of the Δ-t Event Graph is precise and natural yet also to avoid possible over fitting (by choosing Δ-t too large).

Plot of how Temporal Topological Features change with Δ-t (CPE)

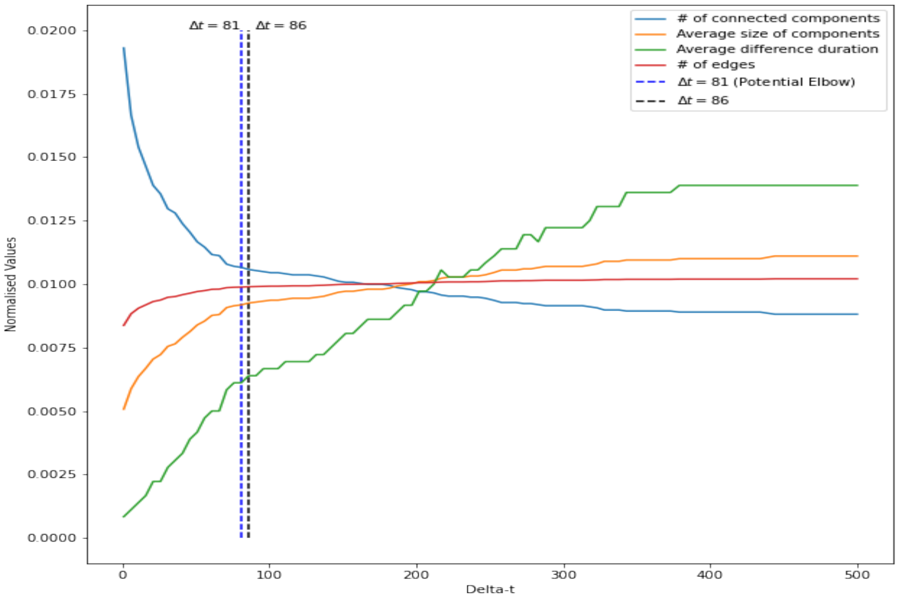


Figure 3.1: Plot of the four temporal and topological features for values of Δ-t in the range [0,500] which have been normalised for comparison. We then search for the average point of maximum curvature over all four features to determine the value of Δ-t, which gives us a value of Δ-t = 81. Shown by the horizontal dashed blue line. We have also included the clinical researched value of Δ-t = 86 for each comparison

Figure 3.1 shows how variations in Δ-t effect the temporal topological structure of the Δ-t Event Graph. We used the Python package 'Kneed' [46], to give a precise way of quantifying the points of maximum curvature (known as the elbow (Convex) or knee (concave)). The point at which the knee/elbow is selected is tunable by setting an internal sensitivity parameter, S. S allows us to adjust how aggressive we want 'Kneed' to be when detecting knee/elbow points, the smaller the value of S, the more aggressive 'kneed' is. Since we wish to detect the most influential knee/elbow point and don't wish to absorb any flat points which may distort the true knee/elbow, we set the parameter S to its lowest value of 1.0.

The average point of maximum curvature (knee/elbow point) of these four temporal and topological measures gave an average value of Δ-t (incubation period of CPE) to be 81 Days.

In fact, considering only the average size of components, the knee point equaled exactly 86 days. This is very exciting as it illustrates how representative our data is in regards to a CPE population and shows how our unsupervised method can be used to find a value of Δ-t which supports the clinical research.

3.1.2 CPE Event Graph Creation

Given the consideration mentioned above we take the value of the parameter Δ-t to be 86-Days and are now in a position to build the Event Graph, [35]. In the Appendix A.1, we have outlined the algorithm created by *Mellor* [33] (available in his Python Package 'eventgraphs' [36]) used to build the Event Graph. The Δ-t Event Graph is a subgraph of the Event Graph where adjacent edges are removed if the Δ-t threshold is not met. Formally, the set of edges of the Δ-t Event Graph are given by $L_{\Delta t} = \{(e_i, e_j) \in | t_j - t_i \leq \Delta t\}$ with the Δ-t Event Graph given by $G_{\Delta t} = (E, L_{\Delta t})$, where E is the set of events.

This decomposes the event sequence into smaller subsequences (restricted by Δ-t threshold), which are represented by the set of weakly-connected components of the Δ-t Event graph, $C_{\Delta t}$. Each component $c \in C_{\Delta t}$ is described by the tuple $(E^c_{\Delta t}, L^c_{\Delta t})$ with $\bigcup_c E^c = E$ and $\bigcup_c L^c_{\Delta t} = L_{\Delta t}$.

	Patients	Events	Duration (Days)	Components
CPE 86-day EG	363	1608	651	106

Table 3.1: Summary statistics for CPE 86-Day Event Graph

Table 3.1 summaries the attributes of the CPE 86-Day event graph. The nodes represent the number of individual patients in the event graph, which is 363. The vertices represent the number of events (contacts between two patients) for which there are 1608. The duration represents the difference between the first contact event and the last, given to be 651 days. The (weakly-connected) components represent sequences of contact events, which in context of this thesis are capturing contact interactions over which transmission can occur, of which there are 106.

3.1.3 COVID-19 Modelling Considerations

Due to the current global pandemic, as a result of COVID-19, there exists extensive research confirming a reliable value for the infectious period of the virus i.e. the number of days someone testing positive with COVID-19 is infectious for. This will be our value of the Δ-t parameter. In this study [61] the infectious period is given to be 10 days. This prolonged incubation time, in combination with pre-symptomatic transmission, is why when tested positive for COVID-19, under UK-law, you are required to self-isolate for 10 days since the start of your symptoms [31].

Following the same unsupervised method described above, we examine how variations in Δ-t affect the temporal topological structure of the Δ-t Event Graph, by again considering the four temporal and topological measures listed above on the event graph. Again for comparison, we normalised and plotted how these changed with variations in Δ-t and using the unsupervised heuristic known as the 'Elbow Method' we attempt to quantify the optimal value of Δ-t.

Plot of how Temporal Topological Features change with Δ-t (COVID-19)

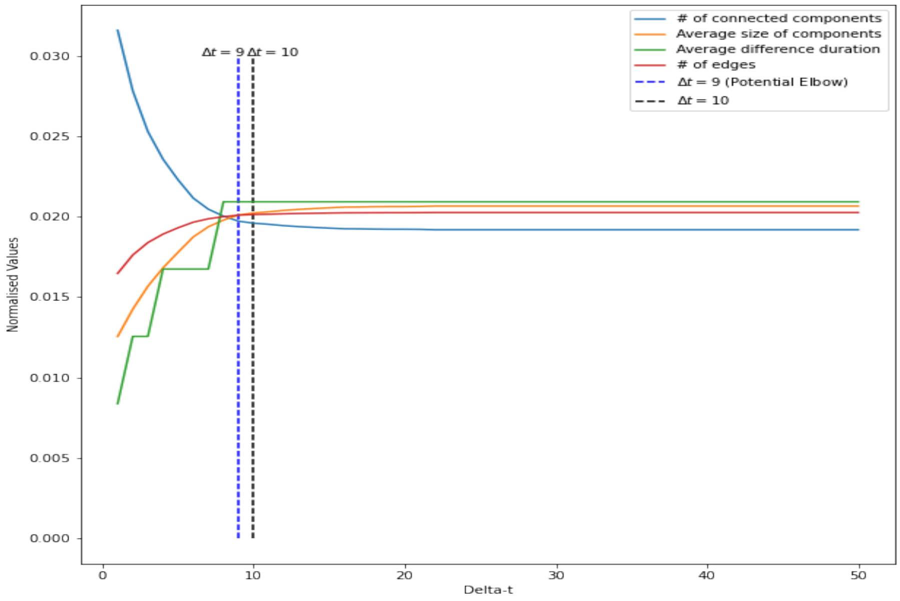


Figure 3.2: Plot of the four temporal and topological features for values of Δ-t in the range [0,50] which have been normalised for comparison. We then search for the average point of maximum curvature over all four features to determine the value of Δ-t, which gives us a value of Δ-t = 9. Shown by the horizontal dashed blue line. We have also included the clinical researched value of Δ-t = 10 for each comparison

As above, we use the 'Kneed' package [46] to quantify the point of maximum curvature. Figure 3.2 shows how variations in Δ-t affect the temporal topological structure of the Δ-t Event Graph. The average point of maximum curvature of the four temporal and topological measures gave a value of Δ-t (prolonged incubation time, in combination with pre-symptomatic transmission period) to be 9 days, which aligns with the clinical research in [61].

Additionally in our modelling of COVID-19 we consider an incubation period (pre-symptoms), where the contacts of a patient 14 days prior to the patients first positive test could have been infected with COVID-19 [30] [15]. As such, we will take the Δ-t parameter to be 24 days, which represents the times when a patient could have been infected ((t-14) day incubation period) or been an infector ((t+10) day infectious period [61]), where t represents the day of first positive test

3.1.4 COVID-19 Event Graph Creation

Given the consideration mentioned above we take the value of the parameter Δ-t to be 24-Days and are now in a position to build the Event Graph, [35]. In the Appendix A.1, I have outlined the Algorithm created by *Mellor* [33] (available in his Python Package 'eventgraphs' [36]) used to build the Event Graph. The Δ-t Event Graph is a subgraph of the Event Graph where adjacent edges are removed if the Δ-t threshold is not met.

As mentioned in 3.1.2 this decomposition breaks the event sequence into smaller subsequences (restricted by Δ-t threshold), which are represented by the set of weakly-connected components of the Δ-t Event graph, $C_{\Delta t}$.

	Patients	Events	Duration (Days)	Components
COVID-19 24-day EG	1672	11146	77	823

Table 3.2: Summary statistics for COVID-19 24-Day Event Graph

Table 3.2 summaries the attributes of the COVID-19 24-Day event graph. The Nodes represent the number of individual patients in the event graph, which is 1672. The Vertices represent the number of events (contacts between two patients) for which there are 11146. The duration represents the difference between the first contact event and the last, given to be 77 Days. The components represent sequences of events, which in context of this thesis are capturing contact interactions over which transmission can occur, of which there are 823.

3.2 Feature Extraction

The components of our event graphs capture sequences of events which occur in relatively close proximity ($0 \leq \tau \leq \Delta$-t) and share one or more nodes. In the context of this thesis, these weakly-connected components represent contact sequences, likely to capture the route of possible transmission/outbreaks of each disease. Importantly these components do not represent the outbreaks themselves, but instead the sequences of contact events which could have lead to disease transmission.

Our goal will be to analyse these (weakly-connected) components to gain an insight and understanding of how these contact sequences could possibly lead to the transmission of the respective diseases within the Imperial College NHS Trust, with the aim to provide advice on and insight into possible route to prevent future outbreaks.

To do this we consider a number of scale-invariant features, both temporal and topological, which will best describe the data and allow us to gain a more realistic insight in to the underlying behaviour and structure of these components. We consider scale-invariant (independent of the component size) features only as naturally this allows for fair comparison. In the rest of this section we describe the features used in this study.

3.2.1 Motifs

In the set-up of the Δ-t Event graph an additional characteristic is that each edge is also associated with a two-event temporal motif which describes the topological relationship between the nodes in each event [27]. Temporal motifs are repeatedly observed structures of interaction across time. Temporal motifs are perhaps the most insightful feature into the behavioural characteristics of temporal networks and have been widely investigated in other studies [42] [59].

We consider temporal motifs as defined by *Kovanen et al.* [27]. In particular we consider only two-event motifs, of which there are six [27]. Mathematically, we define the function $\mu : E \to M$ where E is the set of Edges and

$$M = \{ABAB, ABBA, ABAC, ABCA, ABBC, ABCB\}$$

is the set of two event motifs as defined in *Mellor* [33], which describe the relative positions of the nodes between events. See Figure 3.3.

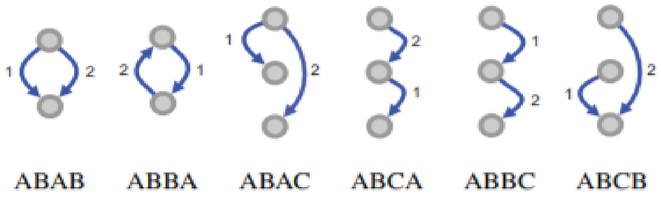


Figure 3.3: All possible two-event motifs. Events are labelled with the order which they occur. Nodes (vertices) represent individual patients, not events. This is only true when considering motif types. Figure taken from *Mellor* [33].

The letters A, B, C represent the vertices (unique patients). The events are represented by the source and target, i.e. AB is the event between the source A and target B. In Figure 3.3, the grey vertices represent patients and blue lines represent events. Events are labelled with the order in which they occur. This representation is only used when dealing with motifs. For every other feature and consideration made, a vertex represents an event and vertices are connected if they share a patient in common. This is a subtle but important difference for our intuition and the reader must keep this in mind.

These different motif types are capturing different structures of transmission in our network. They provide a key intuition into understanding the contact structure over which transmission occurs within the Trust. Below we outline, the interpretation behind each motif type with respect to the context of this thesis. Then, in section 4.5, we gain an insight into their decomposition with the event graph components.

Motif Type: ABAB

Motif type 'ABAB' represents repeated interactions between patients with the same direction. Such repeated interactions are largely the result of how patients are being manged in the trust, rather than an infectious attribute of the disease itself. The motif represents patients who are kept on the same wards and as a result have repeated interactions and therefore have multiple opportunities for the same direction of transmission. These interactions can occur between patients over consecutive days until either a patient is moved to a different ward, isolated or discharged from the Trust.

This motif type is the most common (see Figure 3.4) because when a patient is admitted to hospital they are usually placed on the same specialist ward to be treated, rather than transferred to multiple wards. This is why we see a high prevalence of this motif.

Motif Type: ABAC

The motif type 'ABAC' represents opportunities for a superspreading event where one source is able to infect multiple targets. This is particularly important when it comes to COVID-19 because it's attributes possess these superspreader dynamics [55]. In a recent study conducted by *Illingworth et al.* [22], during the first wave of the COVID-19 epidemic at Cambridge University Hospitals NHS Foundation Trust they found that 21% of individuals caused 80% of transmission events. Similar results were obtained in a study over a larger more general population [1].

Motif Type: ABCB

Motif type 'ABCB' represents the case when one patient has had multiple opportunities to be infected. This is the inverse of motif type 'ABAC'. Comparing this to (non-temporal) graph theory for infectious diseases, this is equivalent to the multiple routes a sink could be infected. (In infectious disease modelling, there are two types of patients. A source which spreads the disease and a sink which is susceptible to being infected). Therefore, we could say that (sink) patients which are involved in this motif type are more likely to possess the disease, just because they have had more possible routes to be infected, than a patient who is say in a transmission chain where there is only one consecutive route.

Motif Type: ABBC

Motif type 'ABBC' represents transmission chains, where a source patient interacts with a target patient, then that target patient becomes a source patient who then interacts with another (target) patient, forming a consecutive chain between patients. (Imagine a relay, where the baton is the disease.)

The motif 'ABBC' doesn't occur that often in our network because of hospital dynamics. In general, most patient when admitted to hospital remain on the same ward where they receive specialist treatment. In order for this motif type to be present, a patient must be admitted to a ward, interact with another patient on that ward, then that patient would have to be transferred to a different ward (or a new admission arrive) the following day. This is a very uncommon scenario, see Figure 3.4. However the wards where this might happen are most likely surgical or ITU wards as they tend to have quick turn over of patients.

Motif Type: ABBA

The motif type 'ABBA' represents a reciprocal event, as an event from A to B is then followed by the reciprocal event B to A. In terms of disease transmission this motif type is of no use since it makes it impossible to determine who infected who as either patient A or B could have been the source. Because of this, we omit the motif type 'ABBA' from our analysis.

Motif Type: ABCA

Motif type 'ABCA' represents a non-sequential contact event between three patients. This is a non-transversable chain of transmission, where there is no route for the disease to transfer between patients B and C due to the time ordering of contacts. This is the least common motif type in our network. The presence of this motif is not a concern because we are clustering all the structures of patient movements within the trust, therefore most of these structures (motifs) will have the pathogen moving over them, allowing disease transmission, however type 'ABCA' will not transmit the disease from B to C because of the time order of events. i.e. if patient A interacts with patient B at time stamp 1, then patient C interacts with patient A at a later time, time stamp 2, then there is no possible route of transmission to Patient C from Patient B. However there is possible route for transmission between A and B and C and A, but no direct transmission from A to B to C.

This is another reason why representing temporal contact networks as an event graph is very natural, because the temporal information restricting the disease from spreading over these motifs would not be picked up if we were considering a non-temporal graph and hence could lead to incorrect infomation on disease transmission. We therefore can distinguish between motifs which allow for disease transmission and those that do not, where as in non-temporal network this distinction cannot be made, due to a lack of temporal ordering.

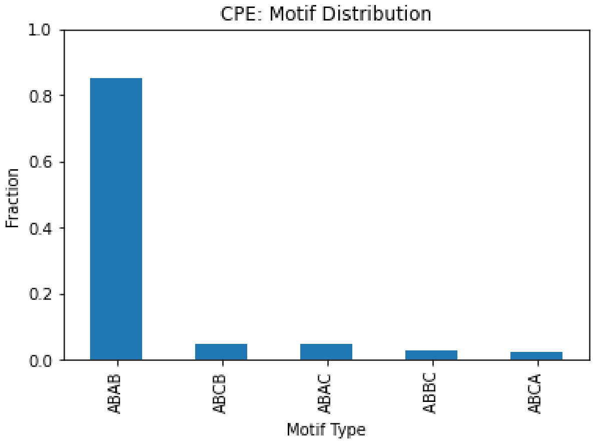


Figure 3.4: Motif distribution for CPE 86-Day Event Graph.

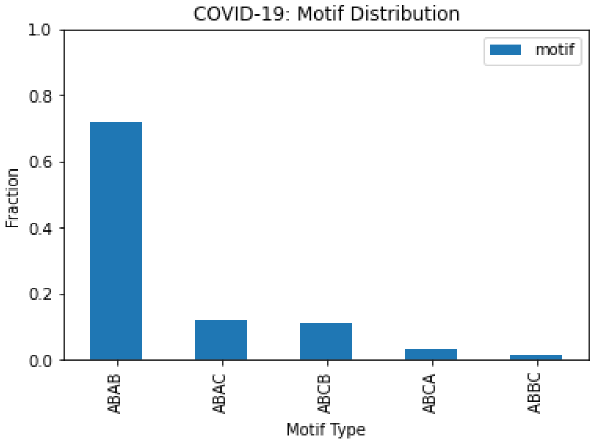


Figure 3.5: Motif distribution for COVID-19 24-Day Event Graph.

Motif Distribution

We create a motif distribution over each outbreak in the event graph by considering the number of each motif type present in a outbreak as a fraction of all motifs present in the event graph.

This allows us to rank motif prevalence and therefore enables us to understand which structures of transmission are more frequent in a contact sequence. For motif prevalence to be meaningful it requires comparison with a suitable null model [27]. Since temporal networks do not have a configuration model equivalent, we will use an ensemble of time-shuffled data as a reference. See Section 4.7 for further details and comparison.

Another feature we wish to extract is the diversity of a contact sequence in terms of the motifs present, which can be derived through the motif entropy defined as:

$$S_{Motif} = -\sum_{m \in M} p_m log_2(p_m)$$

here M is the set of all possible motifs, and p_m is the probability of observing motif type m. This has the desired property that $S_{motif} = 0$ when only one type of motif is present, and takes a maximal value of $log_2(24) \approx 4.58$ when all motif types are equally likely.

Inter-Event Times

The next temporal feature we consider is the distribution of inter-event times along all edges of the Δ-t Event graph. Inter-event times (IETs) have been identified as an import feature of temporal networks, which can dramatically change the properties of a spreading process, [29] and has been the focus of many studies including this study conducted under *Stehlé et al.* [50] where they proposed a rapid communication modeling framework for the spread of information, through interacting social agents at a conference, based on the distribution of inter-event times.

Since the inter-event time distribution is dependent on the duration of the outbreaks, it is not a scale-invariant feature. Therefore, to allow for comparison between outbreaks we instead consider the entropy with the aim to classify the diversity of each outbreaks with respect to the inter-event times. The inter-event time is a continuous variable we therefore first need to discretise. Let I be the set of intervals that partition the set of inter-event times and p_i be the probability of an inter-event time being in interval $i \in I$. Normalising by the maximum entropy gives the inter-event time entropy:

$$S_{IET} = -\frac{1}{log_2 \left|I\right|} \sum_{i \in I} p_i log_2(p_i)$$

The inter-event time entropy has the property of being zero for periodic events and is maximal for a uniform distribution of inter-event times. [33].

Activity

The final temporal feature we consider is known as the Activity of a contact sequence denoted by λ which is defined as the number of events per unit of time. To normalise the activity feature we apply the transformation,

$$\hat{\lambda} = 1 - \exp^{-\lambda}$$

which takes values in $[0, 1)$

The activity allows us to capture the burstiness of the conact sequence. Larger bursts, given the Δ-t parameter is restricting the event graph to capture plausible disease transmission routes, would present more opportunities for large outbreaks/transmission paths.

3.2.2 Topological Features

For the topological features we consider the static graph aggregation of a component. Which is given by the adjacency matrix defined as:

$$A^{(c)}_{uv} = \begin{cases} 1 & if \; \exists t \; s.t \; (u, v, t) \in E^c \\ 0 & \quad otherwise \end{cases}$$

That is, an edge is present in the aggregated graph if an event occurred along that edge at any point in time during the outbreak [40]. This will of course remove any temporal information about the outbreak. However there exists extensive literature on the structural features of static (non-temporal) graph and how to characterise them [40].

We consider four features which are well defined in the literature, namely the edge density, the clustering coefficient, the degree assortativity and the edge reciprocity. We also consider a further feature named the 'Degree Imbalance' as defined in *Mellor* [33].

Edge Density

The edge density is the number of edges present in the graph as a fraction of all possible edges. Mathematically this is given by

$$\rho = \frac{1}{N(N-1)} \sum_{i,j} A_{ij}$$

where N is the number of nodes in the graph [6]. The presence of a N^2 term in the denominator indicates that this feature is only scale-independent on the assumption that the network is dense (i.e. node degrees k_i are $\mathcal{O}(N)$) Since we do not wish to make that assumption, the edge density will only be used as a descriptor and not a feature.

Clustering Coefficient

An important property of any network is the clustering coefficient which measures the degree of transitivity of a graph (local density). For a social network it quantifies the likelihood that the friend of your friend is also your friend. It is defined as the probability that two neighbours of a node will also be neighbours to each other and is given by the ratio of the number of closed triplets of nodes to the number of connected triplets of nodes [54].

Importantly for epidemiology, the clustering coefficient has been shown to effect the ability of a disease to spread [24] [11]. Any change in the behaviour or state of an individual will most likely be dependent on the state of its neighbours (its contacts) i.e. a susceptible individual surrounded by infectious neighbours is likely to become infected when a contact event occurs.

Degree Imbalance

The number of edges emanating from a node is called the degree and indicates the number of possible contacts that can lead to disease transmission to or from an individual. How the degree of each node is distributed is vital to the ability of a disease to spread through a population [37]. As a reminder to the reader, since we are now dealing with a topological feature of the (non-temporal) aggregated graph, and not the event graph, the number of in/out degrees is no longer restricted to a maximum of two.

The degree imbalance is a feature defined in *Mellor* [33], which measures the average difference between the degrees of connected nodes. Let $\alpha, \beta \in in, out$ index the degree type, and let s_i^α and t_i^β be the α- and β-degree of the source and target node for edge i. The degree imbalance is given by:

$$\mu_{(\alpha,\beta)} = \frac{m^{-1} \sum_i (s_i^\alpha - t_i^\beta)}{max_i |s_i^\alpha - t_i^\beta|}$$

where m is the number of edges in the graph, and we sum over all possible edges. We normalise by the maximum difference between node degrees. The degree imbalance takes values in $[-1, +1]$ with a value of $\mu_{(\alpha,\beta)} = \pm 1$ indicating that for all edges the α-degree of the source is greater/less (resp.) than the β-degree of the target, and this difference is the same for all edges.

We use this feature to assess how ‘hub-like’ the contact sequence is. Events in a contact sequence which have a high degree in-out act as ’hubs’ in that they are central to multiple paths of disease transmission.

3.3 Chapter Conclusion

In this chapter, we have outlined the decomposition we used to create our Event Graphs. The construction of an Event graph requires the choice of the parameter value Δ-t which thresholds the time between adjacent events occurring for us to consider them connected. In the context of this thesis, Δ-t represents the maximum time in which a patient can interact with another patient with the possibility of disease transmission to occur over that event. We discussed our method of choosing this value by considering four topological and temporal features and examining how variations in Δ-t affected the temporal topological structure of the Δ-t Event Graph. We quantified the average maximum point of curvature across these features and compared that value to the respective clinical research. We found this method to be incredibly effective at finding the value of Δ-t for Both CPE (Incubation Period) and COVID-19 (Infectious Period), with our chosen values aligning with the clinical research.

In the final part of this chapter, we outlined the temporal and topological features we will use to measure the similarities between the components (contact sequences) of our Event Graphs. Furthermore, we provide our intuition behind these features in terms of their behavioural, compositional and temporal characteristics and relate each feature to the context of this thesis.

Now that we have successfully created an Event Graph representation of the underlying Temporal Network, in the next chapter we seek to understand how these contact sequences could lead to possible routes of disease transmission within the Trust and look to cluster these contact sequences based on the similarities of their features. We aim to gain an insight into the contact structures of these groups of contact sequences with a focus to provide recommendations to the Imperial College NHS Trust on ways to reduce the potential transmission of these diseases throughout the hospital.

Chapter 4

Results: Event Graph Analysis

In this chapter we aim to understand how these contact sequences can capture disease transmission within the Trust. We hope to gain valuable insights into the contact structures and diversity of the contact sequences, so that we may provide recommendations to the trust on ways to prevent future disease transmission. Our analysis is split into two parts, firstly component embedding and clustering then cluster analysis.

Before we cluster our contact sequences (components) of our event graphs based on the similarities of the features described in Chapter 3, we first consider how the events within the twenty largest components of each event graph are dispersed temporally.

4.1 Temporal Barcode

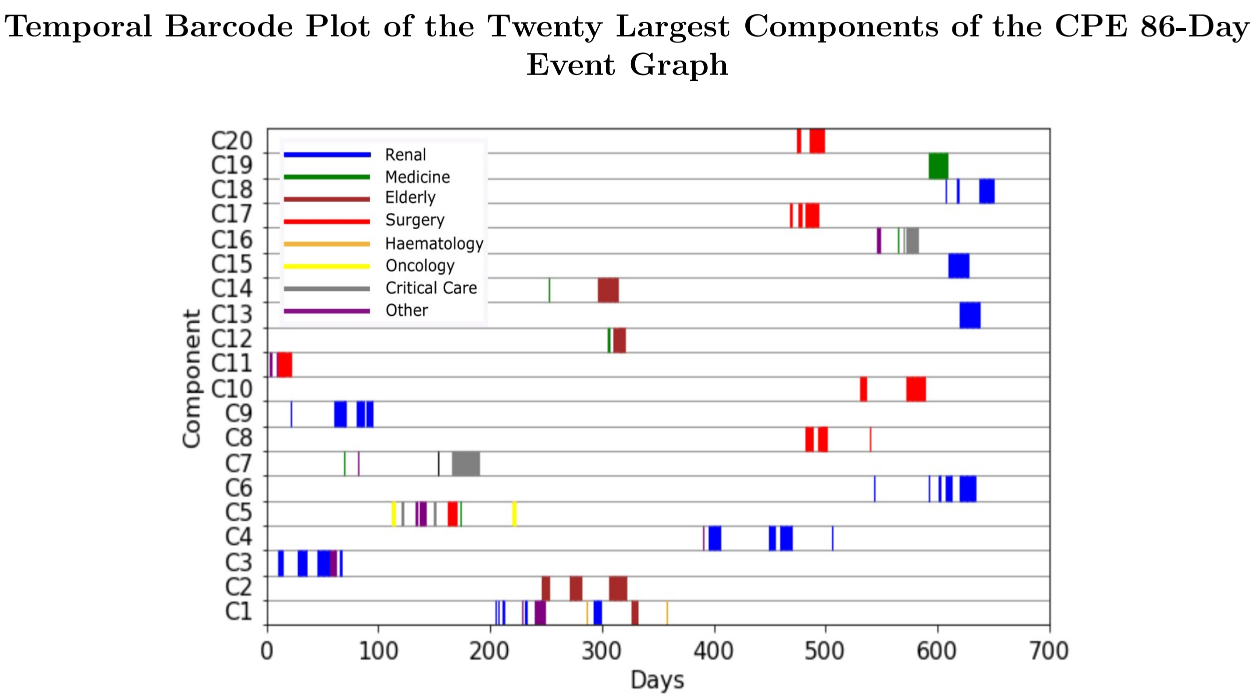
In Figure 4.1 and Figure 4.2 we plot the activities of the twenty largest components over the whole duration of the CPE 86-Day Event Graph and the COVID-19 24-Day Event Graph respectively. Each vertical line represents the time at which an event occurred and we have coloured each event based on the location (ward type) within the Imperial College London NHS Trust.

These plots are incredibly insightful as they allow us to visualise at what times each of the top twenty largest components occur relative to each other and allow us to understand how the locations of these potential transmission pathways (components) are distributed component-wise and relative to the other components. Furthermore, we are able to see the distinct patterns in behaviour in terms of event density (event distribution), component duration, and inter-event time distributions (gaps between events).

4.1.1 Observations from the CPE 86-Day Temporal Barcode

In Figure 4.1 we plot the activity of the twenty largest contact sequences of the CPE 86-Day Event graph, over the whole duration of the data (659 days). Immediately we can see the differences of the temporal activity patterns of these contact sequences; some occur only within the first 100 days and nothing there after (C3, C9, C11) where as others occur over a longer period (C1, C5, C7). Interestingly, none of the 20 largest outbreaks span the entire duration, instead each outbreak is concentrated over a relatively short amount of time and the largest does not last longer than 250 days. We could infer from this that the Trust is able to pick up when patient's are positive and that it is able to do so within 250 days. Contact sequences end when all patients involved are screened, identified with having CPE and then cohorted into an isolation room away from other patients to prevent further contacts. Importantly, for a portion of time they will be infected, not known to the hospital and still interacting with other patients with opportunities for transmission to continue. Therefore under current methods the Trust is able stop a contact sequence and thus the possible further transmission of CPE around the Trust, within 250 days of the first contact event.

We have colour coded the Temporal barcode plots with respect to the ward the event took place on. The most frequent colour is blue which represents Renal wards. This is no surprise as CPE, due to its attributes and placement (i.e. that it is a bacteria that colonises within the gut), is associated with renal symptoms (i.e. in rare cases can cause urine, kidney or bloodstream infections in particularly vulnerable patients). There is also a high proportion of components isolated to Surgery (Red) (C10, C17, C20) which could suggest there is potentially a pool of CPE circulating on these wards. This could be a concern given the variety of patients



Temporal Barcode Plot of the Twenty Largest Components of the CPE 86-Day Event Graph

Figure 4.1: Temporal barcode plot of the twenty largest components of the CPE 86-Day Event Graph. The largest 20 temporal components (by number of events) are plotted (C1 being the largest). Each vertical line represents the time at which an event occurred.

passing through surgical wards, with varying degrees of health and vulnerability. Notably, Elderly care wards (brown) are not well distributed in the top 20 components instead there are single components which Elderly care is concentrated too (C2, C12, C14) which suggests there could also potentially exist a pool of CPE circulating on Elderly care wards, but that they are concentrated i.e. that Elderly patients remain on the Elderly care wards and are not transferred to other wards. They also are concentrated around the 300 day mark. There could be many reason why this is the case. For instance, an influx of new admissions during this period from outside the hospital.

There are clear variations in the inter-event times for the contact sequences (components). C2 for example is made up of three separate but dense periods of activity with the gaps between these periods lasting roughly thirty days. Contact sequences with solid colour block areas (such as C13 and C15) represent times when a contact sequence has a very consistent activity, with consistently small inter-event times. This will be associated with repetitive interactions between two patients on the same ward, as a result of the contact structure 'ABAB'. We can see this by the fact the colour is consistent, wards are not changing and that the inter-event times are consistent. Areas that have bursty behaviour (such as C5) on the other hand, represent inconsistent variable activity and where activity occurs with large variations of inter-event times. This is associated with diverse contact structures, and occurs when patients transfer wards or as the result of a new admission, hence why these areas have varying colours (for further information on these activities see Section 4.5, where we analyse this in more detail).

The key observations from this plot, is that we are seeing the 'Renal', 'Surgery' and 'Elderly' wards consistently throughout the duration, which suggests these wards have the potential to be reservoirs for CPE activity. The majority of the activity remains on these wards, with few contact sequences dispersing across other wards. Another key observation is that the largest contact sequences last no longer than 250 Days, suggesting the Trust is able to end a contact sequence through their current testing and isolating procedure, but that it takes an exceptionally long time to do so.

4.1.2 Observations from the COVID-19 24-Day Temporal Barcode

In Figure 4.2 we plot the activity of the twenty largest outbreaks of the 24-Day Event Graph, over the whole duration of the data.

The temporal activity of the largest twenty components for COVID-19 24-Day Event Graph is very consistent for most of the contact sequences (components), as their inter-event time distribution is very evenly spread with most inter-event times lasting a day. For example, C5 outbreak initially starts on the 41st day and remains constantly active (with an inter-event time distribution uniform a day) until its end at 72 days. Similar to what was mentioned in the analysis of CPE, this consistent activity is related to the motif type 'ABAB' (contact structure) where the same two patients consistently interact over consecutive days on the same ward. However, in contrast to CPE, we see that most components (contact sequences) contain more than one ward. In fact, if we refer back to Figure 3.5 we can see this is as a result of a wider motif distribution, with a lower fraction of 'ABAB' contact structure. As mentioned, motif types, other than ABAB, occur over events where a patient has been moved throughout the hospital (See Section 4.5 for more details via Representative examples) and tend to have large inter-event times been consecutive events in the contact sequence. This is what we are seeing in the majority of the largest components consider C8, we are seeing very bursty behaviour (large inter-event times) across multiple wards with a long duration.

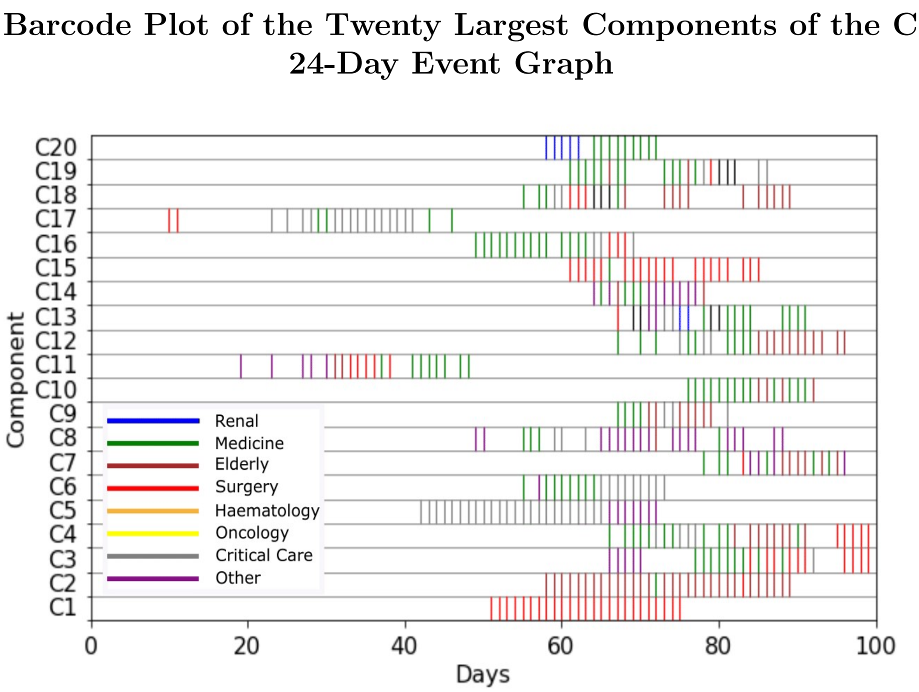


Figure 4.2: Temporal barcode plot of the twenty largest components of the COVID-19 24-Day Event Graph. The largest 20 temporal components (by number of events) are plotted (C1 being the largest). Each vertical line represents the time at which an event occurred.

Surgical wards (red) are present in most of the largest 20 components of the COVID-19 24-Day Event Graph with a couple, including the largest, making up the whole component (C1, C15). From our knowledge of other studies currently being undertaken at Imperial College London, related to COVID-19 in the Imperial College London NHS Trust, which suggest that there has been a real issue with circulation of outbreaks of COVID-19 on surgical wards. Which as we can see in our own analysis is consistent with what these other studies currently being undertaken have found. We are seeing different contact sequences occurring within surgical wards across different and the same times, which suggest to us that there is potentially a pool of COVID-19 circulating on the wards. Medicine is also highly prevalent in these largest components. This makes sense as Medicine is the Trust's, 'general' ward and so many patients will pass through these wards during their admission to the Trust for general treatment, then possibly to more specialist wards such as renal or respiratory for specialist treatment. Hence due to the hub-like nature of the Medicine ward, we would expect to see this prevalence and infer that it too could also act as a pool for COVID-19 transmission. We are also picking up components (C4, C17) where both critical care (grey) and respiratory (black) are present together, which would make sense when considering the nature of COVID-19 [14] i.e. that predominantly affects the respiratory system and can lead to severe respiratory failure, which those with respiratory illnesses would be especially vulnerable to, which would require treatment in critical care wards (ITU).

The final comment to make is regarding the duration of these contact sequences. None of the largest components last more than 40 days (C17 has the longest duration), which illustrates the time taken for the Trust to effectively capture those that test positive and to isolate them.

To conclude this section, through the analysis of the Temporal barcode plot, we found evidence to suggest the possible pooling of COVID-19 on the 'Surgery' and 'Medicine' wards. We also see evidence of multiple ward locations within these largest contact sequences, which is worrying as it could potentially lead to extensive dispersion of the virus over these contact events and thus become more difficult for the Trust to control.

The component barcode plots portray only the duration, activity, inter-event times, and number of events in each component. Therefore to gain a greater understanding and insight into the temporal and topological dynamics of the components, we consider component embedding and clustering.

4.2 Component Embedding and Clustering

We create a feature vector x_c for each component of our Δ-t Event Graph using the temporal and topological features defined in Section 3.2. Formally we define an encoding function $f : C_{\Delta t} \to \mathbb{R}^d$ which maps a component into the d-dimensional vector space. In total we have twelve different scale-invariant features. We normalise our feature space by re-scaling to unit length, i.e. $x^c = x_c/|x_c|$.

We wish to cluster components (contact sequences) of each event graph respectively in order to gain an insight into possible structures of transmission that are similar topologically, temporally and compositionally. This will allow us to see if there are certain structures that are more prevalent to disease transmission and more prevalent within the trust.

To cluster the components we first need to define a suitable distance function. We choose to use the Euclidean distance between components which gives us a pairwise distance matrix between all components from which we can construct a hierarchical clustering via Ward's method [53]. To ensure we choose the optimal number of clusters we employ the silhouette Score [45], Davies Bouldin Score [8] and the Calinski Harabasz score [5]. Which are well defined within the literature and an effective way of choosing the optimal number of clusters, by considering the mode of optimal values over all three scores.

All three scores for both CPE and COVID-19 Event Graphs are shown in Appendix A.2. For both event graphs, CPE shown in Figure A.1 and COVID-19 shown in Figure A.2 there is initially a large variation in scores between the number of clusters equaling two and three. This can be interpreted as under fitting and results in a large variance over these numbers of clusters. We therefore base our choice of the optimal number of clusters from three as this is where the scores start to stabilise around a mean point and the variance is much tighter.

The choice of the number of clusters is based on the mode of the best scores over all three indexes, which for the CPE 86-Day event graph is six and for the COVID-19 24-day event graph is five.

4.3 Principle Component Analysis

We can systematically investigate which are the most important features by applying principal component analysis (PCA). In PCA, we aim to reduce the dimensionality of the feature space in such a way that the information in the reduced dimensional feature space is as close to the information in the original feature space.

PCA works by computing a new set of variables called the Principal Components which are the Eigenvectors of the co-variance matrix for the components of the event graph. Considered together, the new variables represent approximately the same amount of information and the same amount of total variance as the original features [23]. Crucially, the total variance remains the same. However, it is redistributed among the Principle components (new variables) in the most 'unequal' way: the first variable not only explains the most variance among the new variables, but the most variance a single variable can possibly explain. Whereas the last Principle Component explain the least variance any variable can explain.

This is important because it allows us to inspect which Principal Components can express the most information in our Feature Space and can help us understand the most important features representing our event graph. Thus understanding which underlying structures are most prevalent and as such, the most important to potential disease transmission within the Imperial College NHS Trust.

4.3.1 Principle Components Analysis for CPE

In Figure 4.3, we plot the cumulative and individual explained variances against the Principle Components for the CPE 86-Day Event Graph. The first Principle Component explains the most variance, 69%, the second Principle Component explains 13% of the variance and the third Principle Component explains 8%. The total explained variance between these first three Principle components is 91%.

In Table 4.1 we examine the top three feature contributions to the first three Principle Components for the CPE 86-Day Event Graph. The most important feature, which describes the information in the event graph the best, (top ranked in first Principle Component) is the Motif Entropy. This is no surprise, as the motif entropy is a measure of the distribution (diversity) of a component in terms of the motifs present, since motifs are the best feature at describing the behavioural characteristics of a temporal network [42], the motif entropy measures how these different event structures (motifs) of possible disease transmission are distributed within a component, thus allowing us to gain an insight into how a contact sequence allows for possible disease

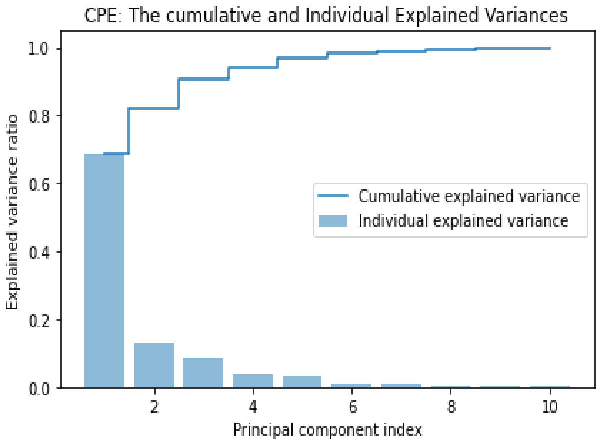


Figure 4.3: Plot of the Cumulative and individual explained variances against the principle components for the CPE 86-Day Event Graph.

transmission. The second most informative feature is the Activity, which is a measure of the burstiness of a component. The Activity is a further indicator of the diversity of a component, it is a measure of the number of potential opportunities for disease transmission within a component. The more bursty a component is the more likely a contact sequence would present opportunities for large and diverse transmission. Finally the third most informative feature, is the motif type 'ABAB'. This indicates that there must be a high contingency for this motif type, which is confirmed in figure 3.4. Intuitively this makes sense, because this motif represents repeated interaction events as a result of patients remaining on the same wards.

	Component 1 (69%)	Component 2 (13%)	Component 3 (8%)
1^{st}	Motif Entropy	Out-In (Imbalance)	Clustering Coefficient
2^{nd}	Activity	ABAB	Activity
3^{rd}	ABAB	Out-Out (Imbalance)	Out-Out (Imbalance)

Table 4.1: Top three feature contributions to the first three Principle Components for CPE 86-Day Event Graph, ranked by magnitude.

4.3.2 Principle Components Analysis for COVID-19

In Figure 4.4, we plot the cumulative and individual explained variances against the Principle Components for the COVID-19 24-Day Event Graph. The first Principle Component explains the most variance, 74%, the second Principle Component explains 11% of the variance and the third Principle Component explains 6%. The total explained variance between these first three Principle Components is 91%.

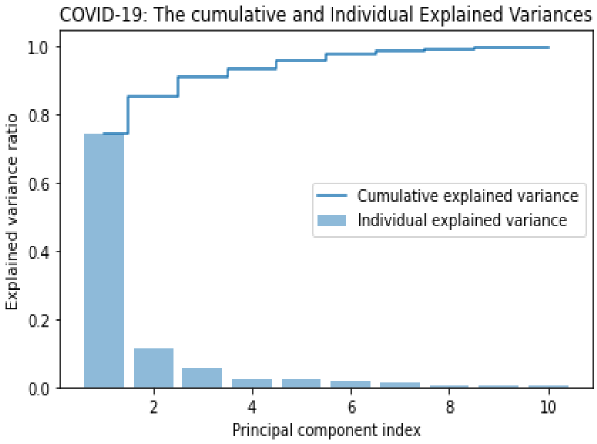


Figure 4.4: Plot of the Cumulative and individual explained variances against the principle components for the COVID-19 24-Day Event Graph.

	Component 1 (74%)	Component 2 (11%)	Component 3 (6%)
1	Motif Entropy	Out-In (Imbalance)	Activity
2	ABAB	ABCB	Motif Entropy
3	Out-Out (Imbalance)	ABAC	Out-In (Imbalance)

Table 4.2: Top three feature contributions to the first three Principle Components for COVID-19 24-Day Event Graph, ranked by magnitude.

In Table 4.2 we examine the top three feature contributions to the first three Principle Components for the COVID-19 24-Day Event Graph. The most important feature which describes the information in the event graph the best, (top ranked in first Principle Component) is the Motif Entropy, which coincides with CPE. Again this is no surprise for the same reasoning as with CPE, that the motif entropy allows us to capture how different event structures (motifs) for possible disease transmission are distributed within a component, thus allowing us to gain an insight into the types of possible disease transmission events. The second most informative feature, is the motif type 'ABAB'. This indicates that there must be a high contingency for this motif type, which is confirmed in figure 3.5. Intuitively this makes sense, because this motif represents repeated interaction events as a result of patients remaining on the same wards. When a patient is admitted to hospital, they generally remain on the same specialist ward for treatment, which is what this motif type suggests. The third most informative feature, is the out-out Imbalance. The out-out Imbalance is a measure of how constricted a disease is to spread for onward spreading. It is quite similar to the superspreader Motif, but here imbalance is computed directly on the events, rather than patients as with motifs.

In Appendix A.4 and Appendix A.5, we have plotted the components for both CPE and COVID-19 in all three Principle Components and overlaid their respective cluster assignments.

4.4 Dendrogram's: Visualising Clusters

A dendrogram is a branching diagram that allows us to represent the relationship of similarities among our clusters of components. We used a bottom-up approach, starting from each component in its own cluster and merging pairs of clusters that are similar as one moves up the hierarchy.

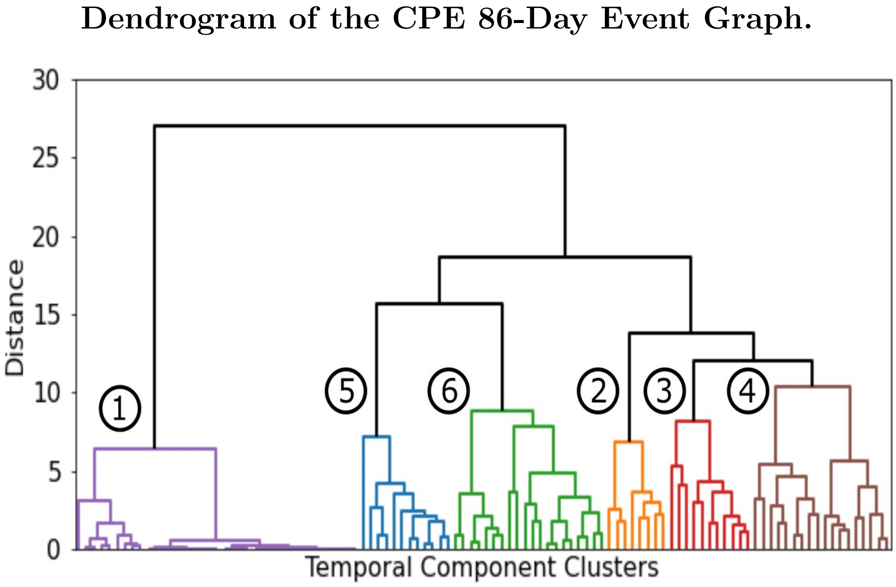


Figure 4.5: Dendrogram plot illustrating the six clusters of the CPE 86-Day Event Graph.

4.4.1 CPE: Dendrogram Analysis

Examining the hierarchy of the CPE clustered outbreaks, shown in Figure 4.5 we can see there are similarities between clusters five (blue) and six (green) and between clusters two (orange), three (red) and four (brown), in particular the most similar clusters are cluster three (red) and cluster four (brown) (illustrated by the lowest join). We can also see the dissimilarity between cluster one (purple) and the other five clusters, with similarity not achieved until the root at a distance of 25. This dendrogram becomes particularly useful when analysing the clusters in Section 4.5.

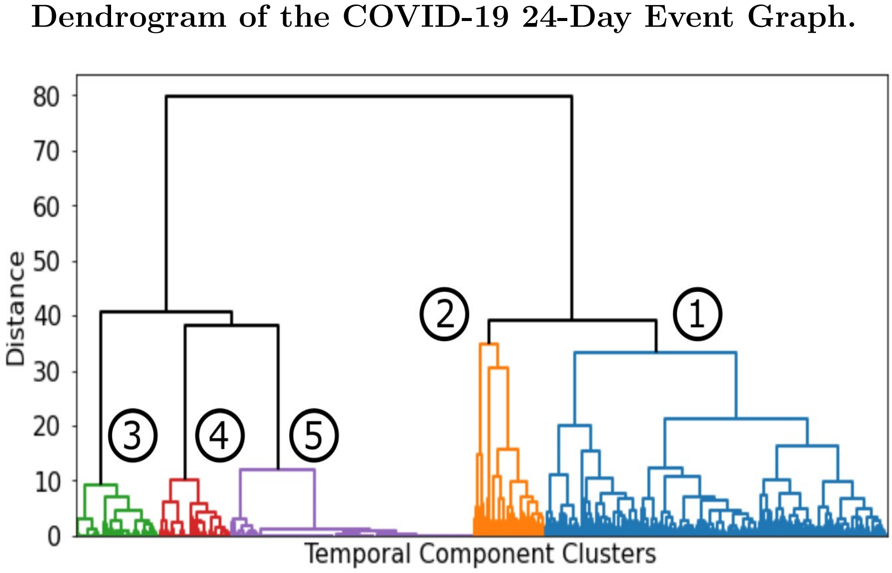


Figure 4.6: Dendrogram plot illustrating the five clusters of the COVID-19 24-Day Event Graph.

4.4.2 COVID-19: Dendrogram Analysis

Examining the hierarchy of the COVID-19 clustered components, shown in Figure 4.6 we can see there are clear similarities between clusters three (green), four (red) and five (purple) and between clusters two (orange) and one (blue). Clusters three (green), four (red) and five (purple) have much lower distances compared to clusters two (orange) and one (blue), which tells us that they are far more similar than clusters two (orange) and one (blue). Again, this dendrogram becomes particularly useful when analysing the clusters in detail in Section 4.5.

4.5 Cluster Analysis

In this section, we analyse each cluster or sets of similar clusters to gain an intuition into how the contact sequences in these clusters are similar and how there contact structures can allow for disease transmission within the Trust. We also pick out representative examples from certain clusters to help the reader visualise the different types of contact sequences we are discovering.

4.5.1 CPE: Cluster Analysis

CPE Feature Vector Averages Across Clusters

Feature	1	2	3	4	5	6
ABAB	1.0000	0.7539	0.6083	0.8171	0.8624	0.7969
ABAC	0.0000	0.0783	0.1169	0.0128	0.0000	0.1347
ABBC	0.0000	0.0761	0.1417	0.0407	0.0000	0.0052
ABCA	0.0000	0.0052	0.0690	0.0773	0.0025	0.0000
ABCB	0.0000	0.0862	0.0638	0.0519	0.1350	0.0630
Activity	1.0367	0.7613	0.2373	0.3812	0.3435	0.4071
Clustering Coefficient	0.0000	0.6681	0.0545	0.0175	0.0000	0.0000
IET Entropy	0.1508	0.6224	1.0492	0.6303	0.6101	0.7192
In-In (Imbalance)	-1.0000	-0.7131	-0.4664	-0.5293	-0.9722	-0.8758
Out-In (Imbalance)	0.0000	-0.1699	0.2015	-0.3276	-0.9722	0.4208
Out-Out (Imbalance)	1.0000	0.6121	0.5184	0.4830	0.9722	0.8741
Motif Entropy	0.0000	2.6768	3.4827	1.9270	1.2493	1.8988
Average Number of Events*	9.5135	16.5000	13.9090	15.6666	12.9166	12.5000
Average Number of Nodes*	2.0000	4.3750	6.1818	4.4444	3.2500	4.0000
Duration*	9.7027	43.7500	66.3636	53.7222	49.5000	43.4500
Edge Density*	0.5000	0.3497	0.1883	0.2521	0.3125	0.2633

*not used as a feature (not scale-invariant).

Table 4.3: CPE feature vector averages across each cluster.

In Table 4.3 we show the (un-normalised) feature vector's averages across each cluster of the CPE 86-Day Event Graph. There are a number of observations to make surrounding the cluster composition.

Cluster 1

The first observation we notice about cluster 1 is that it only consists of motif type 'ABAB' (100 percent). 'ABAB' represents repeated interactions between patients as a result of being kept together on the same ward. Each component in cluster 1 is comprised of two patients (nodes) who remain on the same ward throughout the duration of the contact sequence (component). We are seeing similarities in the way patients are being handled across multiple wards where pairs of patients are interacting repeatedly over consecutive days as a result of remaining together on the same ward. Furthermore the Inter-Event time entropy is very low which is indicative of predictable and systematic behaviour, which concurs with our intuition of repeated (predictable) contact events.

Cluster 1 consists of 37 components out of a total of 106 in the CPE 86-Day Event Graph. Where each component consists of exactly two patients (nodes), which interact over the whole duration of the component in consecutive days on the same wards. The topological structure of these components in cluster 1 are line graphs. This is can be seen by the values of the degree imbalances, $(ii, oi, oo) = (-1, 0, 1)$, and the value of 0.5 for the edge density. We also notice that the duration is very low when compared to the other clusters, this makes sense because each component consists of only two patients and relatively low number of events. Intuitively, the average duration could illustrate the time it takes for the trust to end the contact sequence by discovering all patients that are positive and hence isolate them. We can see here it is very quick, compared to other clusters. These contact sequences could host potential outbreaks but, due to their limited size, would be restricted to just the two patients involved and the one ward location and thus the potential for the pathogen to spread is restricted.

In terms of recommendations to the Trust, this cluster doesn't facilitate the spread of the pathogen well as we have only two interacting patients who interact with themselves only and so the disease spread is contained within these contact events. Secondly the trust is clearly doing well at frequently testing patients on the same wards as the average duration is very low in respect to other clusters implying they are picking up when a patient tests positive quickly and isolating them.

To aid the reader, we provide an Representative example from cluster 1, see in Figure 4.7. Table 4.4, shows two patients interacting in a contact events (vertices) over consecutive days on the renal ward.

Clusters 2, 3 and 4

Clusters 2, 3 and 4 are all very similar, with cluster 3 and 4 being particularly similar. We can see this from the dendrogram see Figure 4.5 and the following analysis.

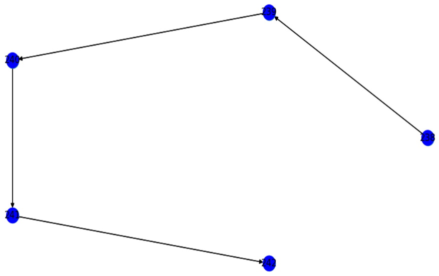


Figure 4.7: Component (index=41) member of cluster 1 of CPE 86-Day Event Graph

Event	Source	Target	Time	Type
238	PtNum_rvketfa70b283rvket	PtNum_rvketcf640db6rvket	67	renal_5
239	PtNum_rvketfa70b283rvket	PtNum_rvketcf640db6rvket	68	renal_5
240	PtNum_rvketfa70b283rvket	PtNum_rvketcf640db6rvket	69	renal_5
241	PtNum_rvketfa70b283rvket	PtNum_rvketcf640db6rvket	70	renal_5
242	PtNum_rvketfa70b283rvket	PtNum_rvketcf640db6rvket	71	renal_5

Table 4.4: CPE cluster 1 representative example (Component 41).

Clusters 2, 3 and 4 all have a motif distribution with each motif type present, with the majority motif being 'ABAB'. These clusters represent the most diverse contact dynamics. We are seeing contact structures which facilitate the transmission of the pathogen extremely well, with diverse spreading dynamics (motifs) across multiple wards. With cluster 3, containing the most contact structures (components) which facilitate disease transmission the best, due to possessing the widest motif distribution.

The combination of having relatively large fractions of motif types 'ABAC' (Superspreader Event), 'ABBC' (Transmission Chain Event) and 'ABCB' (Sink Event) leads to combinations of contact structures that are very diverse and allows for a potentially wide dispersion of CPE transmission, because we are seeing a large number of patients (high average number of nodes) interacting over multiple locations (which we can infer from the motif entropy, as motif types other than ABAB require a change in location in order to occur, by construction). Furthermore, the average duration of the contact sequences in these clusters is very high compared to the other clusters, which is the result of the diversity of the contact sequences which the trust is struggling to control and hence is not picking up when a patient is testing positive quickly enough. This combination of diverse spreading dynamics (motif distribution), high node (patient) count and high ward distribution (motif entropy) results in contact structures which facilitate the spread of the pathogen very effectively and as mentioned, under current methods, the Trust is struggling to control these contacts quick enough to prevent large dispersion throughout the trust (high duration).

The clustering coefficient is particularly high for cluster 2 in respect to all other clusters. As a reminder the clustering coefficient is defined as the likelihood that two neighbours of a node will also be neighbours to each other [54]. It measures the degree of transitivity of a graph (local density). In terms of disease transmission a high clustering coefficient suggest that the contact structures are locally very dense which would imply that the ability for a pathogen to spread over these contact sequences would be very high, which agrees with our intuition of this group of clusters facilitating pathogen spread well.

As a representative example, consider Figure 4.8. Which is an outbreak (index=62) from cluster 3, the most infectious cluster of the 86-Day Event Graph.

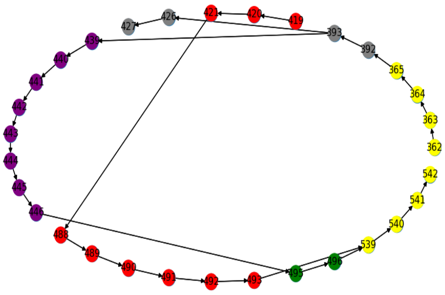


Figure 4.8: Outbreak (index=62) member of cluster 3 of CPE 86-Day Event Graph. Circular layout. Colours represent ward types

Consider the representative example in Figure 4.8 which is a plot of component 62, a member of cluster 3 for the 86-day event graph. This is an example of the diverse behaviour we are seeing in the contact structures (components), within these three clusters. We have coloured contact events (vertices) based on ward types. In this contact structure we see nine individual patients interacting in thirty one separate contact events over five different ward types. Note, there exists far larger and more diverse contact structures within these three clusters. In Table 4.5 we outline the full dataset for this contact sequence including the motif types (other than ABAB with occurs between consecutive timed contact events), the ward types and individual patients. We have included line to break ward types and the double line represents The double

Event	Source	Target	Time	Ward Type
362	PtNum_rvket61f2344arvket	PtNum_rvket253f82dbrvket	112	oncology_2
363	PtNum_rvket61f2344arvket	PtNum_rvket253f82dbrvket	113	oncology_2
364	PtNum_rvket61f2344arvket	PtNum_rvket253f82dbrvket	114	oncology_2
365	PtNum_rvket61f2344arvket	PtNum_rvket253f82dbrvket	115	oncology_2
$ABCA$:	365 → 392			
392	PtNum_rvket9bd252d5rvket	PtNum_rvket61f2344arvket	121	respiratory_2
393	PtNum_rvket9bd252d5rvket	PtNum_rvket61f2344arvket	122	respiratory_2
$ABCB$:	393 → 439 $ABAC$:	393 → 426		
439	PtNum_rvket7d4e3793rvket	PtNum_rvket61f2344arvket	134	cardiology_4
440	PtNum_rvket7d4e3793rvket	PtNum_rvket61f2344arvket	135	cardiology_4
441	PtNum_rvket9c12e357rvket	PtNum_rvket61f2344arvket	137	cardiology_4
442	PtNum_rvket9c12e357rvket	PtNum_rvket61f2344arvket	138	cardiology_4
443	PtNum_rvket9c12e357rvket	PtNum_rvket61f2344arvket	139	cardiology_4
444	PtNum_rvket9c12e357rvket	PtNum_rvket61f2344arvket	140	cardiology_4
445	PtNum_rvket9c12e357rvket	PtNum_rvket61f2344arvket	141	cardiology_4
446	PtNum_rvket9c12e357rvket	PtNum_rvket61f2344arvket	142	cardiology_4
$ABBC$:	446 → 495			
426	PtNum_rvket9bd252d5rvket	PtNum_rvket3c61c1aarvket	150	respiratory_2
427	PtNum_rvket9bd252d5rvket	PtNum_rvket3c61c1aarvket	151	respiratory_2
419	PtNum_rvket41233dc3rvket	PtNum_rvket2a2cab77rvket	162	surgery_12
420	PtNum_rvket41233dc3rvket	PtNum_rvket2a2cab77rvket	163	surgery_12
421	PtNum_rvket41233dc3rvket	PtNum_rvket2a2cab77rvket	164	surgery_12
488	PtNum_rvket41233dc3rvket	PtNum_rvket2a2cab77rvket	165	surgery_12
489	PtNum_rvket41233dc3rvket	PtNum_rvket2a2cab77rvket	166	surgery_12
490	PtNum_rvket41233dc3rvket	PtNum_rvket2a2cab77rvket	167	surgery_12
491	PtNum_rvket41233dc3rvket	PtNum_rvket2a2cab77rvket	168	surgery_12
492	PtNum_rvket41233dc3rvket	PtNum_rvket2a2cab77rvket	169	surgery_12
493	PtNum_rvket41233dc3rvket	PtNum_rvket2a2cab77rvket	170	surgery_12
$ABCB$:	493 → 539			
495	PtNum_rvket61f2344arvket	PtNum_rvket57fffafervket	173	medicine_21
496	PtNum_rvket61f2344arvket	PtNum_rvket57fffafervket	174	medicine_21
$ABAC$:	496 → 539			
539	PtNum_rvket61f2344arvket	PtNum_rvket2a2cab77rvket	220	oncology_2
540	PtNum_rvket61f2344arvket	PtNum_rvket2a2cab77rvket	221	oncology_2
541	PtNum_rvket61f2344arvket	PtNum_rvket2a2cab77rvket	222	oncology_2
542	PtNum_rvket61f2344arvket	PtNum_rvket2a2cab77rvket	223	oncology_2

Table 4.5: CPE cluster 3 representative example (Component 62). Split table based on ward type for easier comparison to Figure 4.8 and included motif types and their source and target events. Note, 'ABAB' motifs connect consecutive timed events and thus have not included (but they exist between each consecutive event). The double line represents where one of the two paths that split at contact event 393 ends.

line represents where one of the two paths that split at contact event 393 ends. From the dataset we can see the distribution of inter-event times and in particular the larger inter-event times where motif types (other than ABAB) occur. There are many possible reasons why these inter-event times are much larger. For example, two patients maybe interacting, then one of the patients could test positive for CPE and be put into isolation, then at a later time a new admission arrives onto the ward which would then start the contact sequence again. Another possibility relating to Renal wards in particular, is the admission of day case patients undergoing dialysis once every week. We can also see the times when patients are picked up on a positive CPE test and hence are put into isolation, these occur when the patient no longer interacts in the contact sequence and are removed. For example, patient 'PtNumrvket253f82dbrvket' no longer interacts in the contact sequence after time 115.

Our recommendations on how to reduce future transmission of CPE throughout the Trust based on the analysis of these three clusters is to increase the frequency of testing on the Hot-Spot wards, which we define as

Definition 4.1 ***Hot-Spot Wards***
We define Hot-Spot Wards as the wards in which a set of clusters are most frequently involved within. These represent the wards which the contact sequences within a cluster pass through the most and thus are potentially hubs for disease transmission.

For the set of clusters 2, 3 and 4 of the CPE 86-Day Event Graph these are, in order of magnitude, 'Renal', 'Surgery', 'Elderly', 'Critical Care' and 'Medicine'.

This should help to pick up potential outbreaks over these contact structures much quicker than current methods. This should help prevent the further potential spread of the pathogen to other patients, by cutting the contact sequences through isolating the positively-tested patients involved. Most importantly, these clusters facilitate the potential spread of the pathogen over the contact sequences the most effectively. Therefore reducing the duration and size (number of patients and number of events) of these contact sequences should rapidly deplete the potential transmission of CPE and reduce its impact with the trust. Out of all clusters of contact sequences, these three clusters are the highest priority to control because of their ability to effectively facilitate pathogen spread.

Clusters 5 and 6

Clusters 5 and 6 are very similar. We can see this from the dendrogram and the following considerations.

Both consist of a majority, over 90 percent, in two motif types. Cluster 5 consists of majority 'ABAB' (repetitive Event) and 'ABCB' (Sink Event) whereas cluster 6 is majority 'ABAB' (repetitive Event) and 'ABAC' (Superspreader Event). Interestingly, we have reciprocal motif types in each cluster. 'ABCB' represents multiple opportunities for transmission to occur with multiple pathways the pathogen could have travelled to a single patient, whereas 'ABAC' represents a Superspreader event, where a single patient causes the possible transmission of the pathogen to multiple other patients. In particular, cluster 5 has no distribution of motif types 'ABAC' (Superspreader Event) or 'ABBC' (Transmisson Chain) and a tiny proportion (0.25%) of 'ABCA' (Non-sequential). Since transmission of the pathogen is not possible over motif types 'ABCA', we can ignore this motif type. Therefore since we are only considering motif type 'ABAB' and 'ABCB' (low motif entropy) for possible disease transmission, what we are seeing in relation to the Trust is a single target patient remaining on the same ward throughout the duration of the component but that the source patients are changing. These changes in source patients occur at various (non-consecutive) times which results in large inter-event time gaps over these 'changing' events. This is why the IET-Entropy is relatively large compared to the motif distribution (ABAB gives low IET-entropy as consecutive events however since we are seeing large IET over the motif type ABCB this weighs on the IET-entropy pushing it up)

This domination from motif type ABCB can also be inferred from the Degree Imbalance values ($(II, OI, OO) = (-0.9722, -0.9722, 0.9722)$), in particular that the average Out-In Imbalance sways heavily negative (if it was just the one motif type ABAB, as in cluster 1, we would expect a value of zero) this is caused by the motif type 'ABCB' where we are seeing higher in-degrees than out-degrees.

We see a similar situation for cluster 6 where, a single source patient (not Target patient as ABAC is reciprocal of ABCB) remaining on the same ward throughout the duration of the component but that the Target patients are changing. However, we do see further distinctions because of the fraction of motif type 'ABCB' coming into play and causing changes in the Source patients. Therefore, the similarities are there but cluster 6 represents a slightly more diverse situation and as a result we see a higher average number of nodes and higher ward distribution, than cluster 5.

They both have zero clustering coefficient. The clustering coefficient measures the degree of transitivity of the graph i.e. how many of a nodes closest neighbours interact with each other. This makes sense under the structures we have mentioned above, that the sources/targets remain the same so there is no opportunity for neighbours to interact.

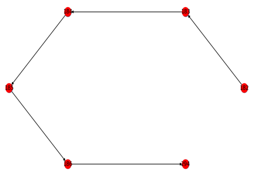


Figure 4.9: Component (index=33) member of cluster 5 of CPE 86-Day Event Graph. Circular layout. Colours represent ward types

Event	Source	Target	Time	Type
182	PtNum_rvket774d8ba5rvket	PtNum_rvket51fd7a15rvket	63	surgery_1
183	PtNum_rvket774d8ba5rvket	PtNum_rvket51fd7a15rvket	64	surgery_1
184	PtNum_rvket774d8ba5rvket	PtNum_rvket51fd7a15rvket	65	surgery_1
185	PtNum_rvket774d8ba5rvket	PtNum_rvket51fd7a15rvket	66	surgery_1
186	PtNum_rvket774d8ba5rvket	PtNum_rvket51fd7a15rvket	67	surgery_1
294	PtNum_rvket89ef548arvket	PtNum_rvket51fd7a15rvket	81	surgery_1

Table 4.6: Component (index=33) data. Member of cluster 5 of CPE 86-Day Event Graph.

Consider the representative example in Figure 4.9 with the corresponding data in Table 4.6. Here we see the target patient remaining the same throughout the whole duration of the component. We also see there is motif type 'ABCB' occurring between nodes 186 and 294. Notice the large jump in time between these two events. There are a number of reasons why this could occur including patient PtNum_rvket774d8ba5rvket testing positive and being isolated and then a new patient PtNum_rvket89ef548arvket admitted to the ward at a time 81.

General Comments and Conclusion on the Cluster Composition of CPE

The vast majority of CPE transmission could potentially be driven by repeat interactions since across all six clusters we are see that the majority motif type present is 'ABAB'. A significant risk factor for CPE infection in hospitals is prolonged admission because they will likely be interacting with the same patients, who maybe positive for CPE, and thus are more likely to be infected.

To conclude, our recommendations to the Trust based on the analysis we have conducted on the clustering of contact sequences and the various contact structures they contain for the CPE 86-Day Event Graph is to prioritise the frequency of CPE testing within the 'Hot Spot' wards. These are, ordered by magnitude, 'Renal', 'Surgery', 'Elderly', 'Critical Care' and 'Medicine'. Our intention with this recommendation is to reduce the diversity and size of contact sequences where the likelihood of CPE spread is increased, by picking up when a patient is positive for CPE much quicker.

our second recommendation is to reduce the number of Superspreader events (ABAC) which results from a single patient interacting with multiple patients, which can happen through frequent ward transfers or new admissions on wards. Thus we recommend, these events are flagged as potentially facilitating the spread of CPE within the Trust and the necessary precautions taken, such as pre-isolation period before new admissions are let onto wards and patients not being transfered from their inital ward unless absolutely necessary.

4.5.2 COVID-19: Cluster Analysis

In Table 4.7 we show the (un-normalised) feature vector's averages across each cluster of the COVID-19 24-Day Event Graph. There are a number of observations to make surrounding the cluster composition.

COVID-19 Feature Vector Averages Across Clusters

Feature	1	2	3	4	5
ABAB	0.5647	0.6099	0.7846	0.7994	1.0000
ABAC	0.1858	0.1258	0.2153	0.0000	0.0000
ABBC	0.0053	0.1386	0.0000	0.0000	0.0000
ABCA	0.0586	0.0287	0.0000	0.0000	0.0000
ABCB	0.1853	0.0968	0.0000	0.2005	0.0000
Activity	1.2252	-2.2723	0.9133	0.9219	1.1508
Clustering Coefficient	0.0063	0.1355	0.0000	0.0000	0.0000
IET Entropy	1.0149	0.9466	0.4762	0.4769	0.0793
In-In (Imbalance)	-0.7237	-0.5726	-1.0000	-1.0000	-1.0000
Out-In (Imbalance)	-0.0077	0.0827	1.0000	-1.0000	0.0000
Out-Out (Imbalance)	0.7346	0.6056	1.0000	1.0000	1.0000
Motif Entropy	3.2205	3.2391	1.6920	1.6257	0.0000
Average Number of Events*	11.9479	10.3750	6.6000	6.8082	6.6356
Average Number of Nodes*	6.3670	5.3333	3.1294	3.0547	2.0000
Duration*	9.3554	10.2916	7.5176	7.7534	5.8906
Edge Density*	0.1889	0.2406	0.3225	0.3287	0.5000

Table 4.7: COVID-19 feature vector averages across each cluster.

Cluster's 1 and 2

Clusters 1 and 2 represent contact sequences that are particularly diverse, represented through the motif distribution over the clusters. We are seeing contact structures (components) which facilitate the transmission of the virus very effectively, with diverse spreading dynamics (motifs) across multiple wards. With cluster 1 containing the most infectious contact structures due to having the widest distribution of motif types.

Cluster 1 has the lowest percent of motif type 'ABAB' only 56 percent, allowing for a high percentage of both motif type 'ABAC' (Superspreader Event) at 18.58% and 'ABCB' (Sink Event) at 18.53%. Both ABAC (Superspreader) and ABCB (Sink event) are the most infective contact structures (motif types) with the highest likelihood of disease transmission over these events because on the one hand, you have Superspreader events meaning one patient is potentially spreading the virus to multiple other patients whilst on the other hand, you have a sink event which encourage the likelihood of infection as a result of multiple contact sequences reaching the same patient. This combination means that Cluster 1 contains contact structures which potentially facilitate the transmission of the virus very effectively. This conclusion can be inferred through the average number of nodes and events, which are the highest in cluster 1 with respect to all other clusters. Hence we are seeing large numbers of patients interacting over multiple wards, which is very concerning when considering the infection rate of COVID-19 [58]. What's more, Cluster 1 contains the most components at 346 out of a total of 823 components in the 24-Day Event Graph.

Cluster 2 contains contact sequences which facilitate the spread of COVID-19 as effectively as cluster 1, since it contains a high proportion of motif type 'ABAC' at 12.58% (Superspreader), 'ABCB' at 9.68% (Sink Event) and 'ABBC' at 13.86% (transmission Chain). The presence of motif type 'ABBC' is a concern since it represents the possible transmission of the virus through chain like behaviour. Where patients are potentially transmitting the disease across patients in a consecutive manner. Within the Trust this contact structure (motif) is only possible due to either movements between wards, where a patient interacts with another patient and that patient is transferred to another ward where they transfer it to a further patient or the through the admission of a new patient onto a ward. Although this contact structure is not potentially as infectious as a superspreader event, it is an added element which adds to the likelihood of potential disease transmission over these contact sequences in this cluster. Furthermore, having a relatively large percentage of these contact structures indicates the current underlying management of patient movements in the trust allows for these contact sequences to be prevalent.

Cluster 2 contains 72 components out of a total of 823 in the 24-Day Event Graph. Similar to cluster 1 we are seeing large numbers of patients (high average node count) interacting in many events (high average event count) over multiple wards (high motif entropy). Cluster 2 also holds the highest average duration which implies the time it takes for the trust to end the outbreak by discovering all patients that are positive and then isolate them (ending their contribution to the component), is longer than it takes for the other clusters. This is most likely because of the diversity of the contact sequences which are harder to contain.

Our recommendations to the trust, based on the analysis of these two clusters is to increase the frequency of testing on the Hot-Spot wards, which for this set of clusters COVID-19 24-Day Event Graph are, in order of magnitude, ’Surgery’, ’Medicine’, ’Critical Care’, ’Other’ and ’Respiratory’. Note, the ward type ’Other’, includes Cancer and Cardiology. This should help to pick up potential outbreaks over these contact structures much quicker than current methods which should help prevent the further potential spread of the virus to other patients, by cutting the contact sequences through isolating the positively-tested patients involved. Most importantly, these clusters facilitate the potential spread of the pathogen over the contact sequences the most effectively so reducing the duration and size (number of patients and number of events) of these contact sequences should rapidly deplete the potential transmission of COVID-19 and reduce its impact with the trust. Out of all clusters of contact sequences, these two clusters are the highest priority to control because of their ability to effectively facilitate viral spread.

Cluster’s 3, 4 and 5

Clusters 3, 4 and 5 are very similar. We can see this from the dendrogram see Figure 4.6 and the following considerations.

Cluster’s 3, 4 and 5 are all similar and all dis-similar to cluster’s 1 and 2 because their motif distributions are more concentrated to one or two particular contact structures (motif types). Cluster 3 consists only of motif type ’ABAC’ (Superspreader), excluding ’ABAB’. Cluster 4, on the other hand is only made up of motif type ’ABCB’ (Sink Event), excluding ’ABAB’. These two clusters can be seen as complete reciprocals of each other, since motif types ’ABCB’ (Sink Event) is the reciprocal of ’ABAC’ (Superspreader Event). As a reminder to the reader, ’ABCB’ represents multiple opportunities for viral transmission to occur to a single patient whereas ’ABAC’ represents a Superspreader event, where a single patient causes the possible transmission of the virus to multiple other patients. In Cluster 3 therefore we are seeing a single source patient remaining on the same ward (’ABAB’) throughout the duration of the component but that the target patients are changing. These changes in target patients occur as a result of motif type ’ABAC’ which happen at various (non-consecutive) times, resulting in potentially large inter-event times over these ’switching’ events. In cluster 4 we are seeing the opposite situation. Instead a single target patient remaining on the same ward (’ABAB’) throughout the duration of the component but that the source patients are changing. These changes in source patients occur as a result of motif type ’ABAC’ which happen at various (non-consecutive) times, resulting in potentially large inter-event times over these ’switching’ events.

Cluster 5 however, consists of only ABAB (100%). As a reminder to the reader, ’ABAB’ represents repeated interactions between patients as a result of being kept together on the same ward. Each contact sequence in cluster 5 is comprised of two patients (nodes) who remain on the same wards throughout the duration of the contact sequence. We are seeing similarities across multiple wards (hence why they are clustered), where pair’s of patients are interacting repeatedly over consecutive days as a result of remaining together on the same ward. Furthermore, the Inter-Event time entropy is very low which is indicative of predictable and systematic behaviour, which concurs with our intuition of repeated (predictable) contact events. This type of behaviour is expected within a hospital because when a patient is admitted to hospital, they are put on specialist wards where they remain for the majority (or all) of their treatment. Another notable feature is the clustering coefficient which measures the degree of transitivity of the graph i.e. how many of a nodes closest neighbours interact with each other. For cluster 5 the clustering coefficient is obviously zero as it contains only contact sequences composed of contact structure ’ABAB’ where each of the contact events are isolated to just two patients and therefore no ’neighbours’ exist.

With respect to recommendations to the Trust, these clusters of contact sequences do not facilitate the transmission of COVID-19 as well as clusters 1 and 2. This is because they are composed of a tighter motif distribution, majority ’ABAB’ (repetitive event) with either ’ABAC’ (Superspreader Event) or ’ABCB’ (Sink Event), and therefore the types of contact structures we are seeing are less diverse meaning the contact sequences within these clusters are more restricted and hence the possible transmission of COVID-19 over these contact sequences will be restricted also, compared to those of clusters 1 and 2. In particular, cluster 5 facilitates transmission the least as it is composed of only repetitive events between two interacting patients on the same ward. Therefore the disease would be contained completely. Furthermore, the Trust is doing relatively well at picking up when a patient is positive and then isolating them because the average duration of contact sequences in these clusters is much lower than those in cluster 1 and 2. We therefore suggest that the Trust does everything it can to prevent these two particular types of contact structures from being allowed. We suggest the ward not only flags these events as a major risk factors for disease transmission but also includes precautionary considerations if the event is essential. For example, if a patient needs to be transferred between wards or a patient needs to be admitted to a new ward they could under go a pre-isolation period, before being admitted to the new ward and extra testing should be undertaken to ensure the patient is not infectious.

General Comments on COVID

The contact structures 'ABAC' and 'ABCB' are the most prevalent thoughtout all clusters (other than ABAB). This suggests that the current methods the trust is implementing are perhaps not considering or flagging these contact structures as risky for potential transmission of the virus. Research shows that COVID-19 is a very infectious virus and can last for extended periods of time out of host therefore these contact structure risk encouraging disease transmission throughout the trust [14].

General Comments on Both CPE and COVID

An interesting observation of our clustering analysis is that for both CPE and COVID-19 we are seeing similar cluster compositions. We therefore can group these cluster types into categories based on the transmission potential within the Imperial College London NHS Trust.

Definition 4.2 ***Category 1: Highly Effective Transmission Potential (HETP)***
Clusters which fall into the Highly Effective Transmission Potential (HETP) category contain contact sequences which could potentially facilitate the transmission of a disease very effectively throughout the Trust because they contain contact structures which encourage wide interaction

between patients across multiple wards. These contact sequences are very diverse and hence are very difficult to contain as shown by the high average duration in these clusters.

For CPE, Clusters 2, 3 and 4 all fall into the HETP Category as these contain contact sequence which have the most diverse contact dynamics and therefore encourage the potential transmission of the pathogen. For COVID-19, Clusters 1 and 2 fall into the HETP Category.

CPE cluster 2, 3 and 4 have a combined number of 37 components which is 35% of all components in the Event Graph. COVID-19 cluster 1 and 2 have a combined number of 418 components which is 51% of all components.

Definition 4.3 ***Category 2: Mild Transmission Potential (MTP)***
Clusters which fall into the Mild Transmission Potential (MTP) category contain contact sequences which could potentially facilitate the transmission of a disease throughout the Trust but are far more contained. These Clusters contain contact structures which involve relatively low numbers of patients across a small number of wards. In particular, we saw the emergence of only two contact structures (motif types), ABAB and then either ABAC (Superspreader Event) or ABCB (Sink Event). Clusters in this category, contain contact sequences which are much easier to contain than those in HETP category which we infer from the lower average duration.

For CPE, Clusters 5 and 6 fall into the MTP Category. They contain contact structures made up of almost entirely repetitive interactions (ABAB) and either a sink (ABCB) or superspreader (ABAC) events, meaning the diversity of these contact sequences is relatively low and since they involve a low number of patients and wards, these clusters are easier to contain than clusters within the HETP category, as shown by an average lower duration of the clusters. For COVID-19, both clusters 3 and 4 fall into the MTP Category (with cluster 5 falling into the NTP category, defined next). Notably, for Cluster 3 and 4 they both contained only repetitive interactions (ABAB) and superspreader (ABAC) and sink (ABCB) events respectively, with no other contact structures present.

CPE cluster 5 and 6 have a combined number of 32 components which is 30% of all components in the Event Graph. COVID-19 cluster 3 and 4 have a combined number of 158 components which is 19% of all components.

Definition 4.4 ***Category 2: No Transmission Potential (NTP)***
Clusters which fall into the No Transmission Potential (NTP) category contain contact sequences which cannot potentially facilitate the transmission of a disease throughout the Trust. Instead any transmission over these contact sequences is completely isolated and there exists no likelihood of transmission beyond those involved. Clusters in this category contain contact sequences with are composed of only repetitive events between two patients on the same ward only. Therefore any potential disease transmission remain with these two patients only and there is no risk of spreading beyond these two patients. Eventually one will test positive and be isolated, ending the contact sequence. These contact sequences have the average lowest duration which implies the Trust is picking up when a patient is testing positive very quickly and isolating them.

For CPE Cluster 1 falls into the NTP category, whereas for COVID-19 Cluster 5 falls into this category. CPE cluster 1 has 37 components which is 35% of all components in the Event Graph. COVID-19 cluster 5 has 247 components which is 30% of all components, therefore we are seeing large fraction of contact sequences fall into this category compared to the other two. Our intuition is that more often than not patients in hospitals require specialist treatment and thus remain on the same ward. resulting in repeated interaction, which is what we are seeing in these contact sequences.

We can infer, that the way the Trust manages patients with COVID-19 compared to CPE is very different and done so in a way which potentially increases the risk of disease transmission. Firstly because the percentage of contact structure within the NTP category is lower than that of CPE, and therefore more contact sequences will fall within the other two more transmissible categories. Furthermore, a total of 51% of all components in the COVID-19 24-Day Event Graph fall within the HETP category compared to 35% of all components for the CPE 86-Day Event Graph. Note that these categories are based on the underlying motif distributions of the sets of clusters and if we compare the overall motif distributions of the COVID-19 24-Day Event Graph to that of the CPE 86-Day Event Graphs, see Figures 3.5 and 3.4, we see that the Trust is allowing far more Superspreader (ABAC) and Sink (ABCB) events to occur compared to CPE. This is a management issue within the trust that needs to be sorted.

4.6 Cluster Evolution Over Time

Our clusters are not equal in size nor do they persist uniformly across time. Each contact event (vertex) is associated with a contact sequence (component) which is in turn associated with a cluster. In the following analysis, we consider how these cluster evolve over time as fraction of the entire Event Graphs. We aim to understand which clusters dominate our networks and which clusters do not and for what periods of time.

4.6.1 CPE: Cluster Evolution's Over Time

In Figure 4.10 we plot the size of a cluster over a particular day (for the period [0,100]) as a fraction of all events. The duration of our CPE dataset is 659 days and in Appendix A.3 we have plotted the whole duration in 100 day intervals.

Figure 4.10 which is over the period [0,100] days, displays the most interesting period of dynamics over time out of the whole duration. Interestingly, we see that there are periods in time when only one cluster makes up the entire network for that period. For example cluster 4 (red) for days in [45,48]. If the reader recalls, Cluster 4 contains contact sequences which potentially can facilitate pathogen spread well. Therefore having dominance in this cluster for extended periods of time is worrying for the Trust. Similarly, clusters 2 and 3 also fall into the category of containing contact sequences which potentially can facilitate pathogen spread well. If we look at their evolution (as well as cluster 4's) as a category on the whole, these three clusters are fairly consistent and in relatively high volumes throughout the whole duration, when compared to clusters 1, 5 and 6 which contain contact sequences which do not facilitate pathogen spread as effectively. i.e. we are fairly consistently seeing clusters 2, 3 and 4 as time progresses, compared to clusters 1, 5 and 6. This is a concern because for certain periods of time the only events that are taking place are those which encourage disease transmission.

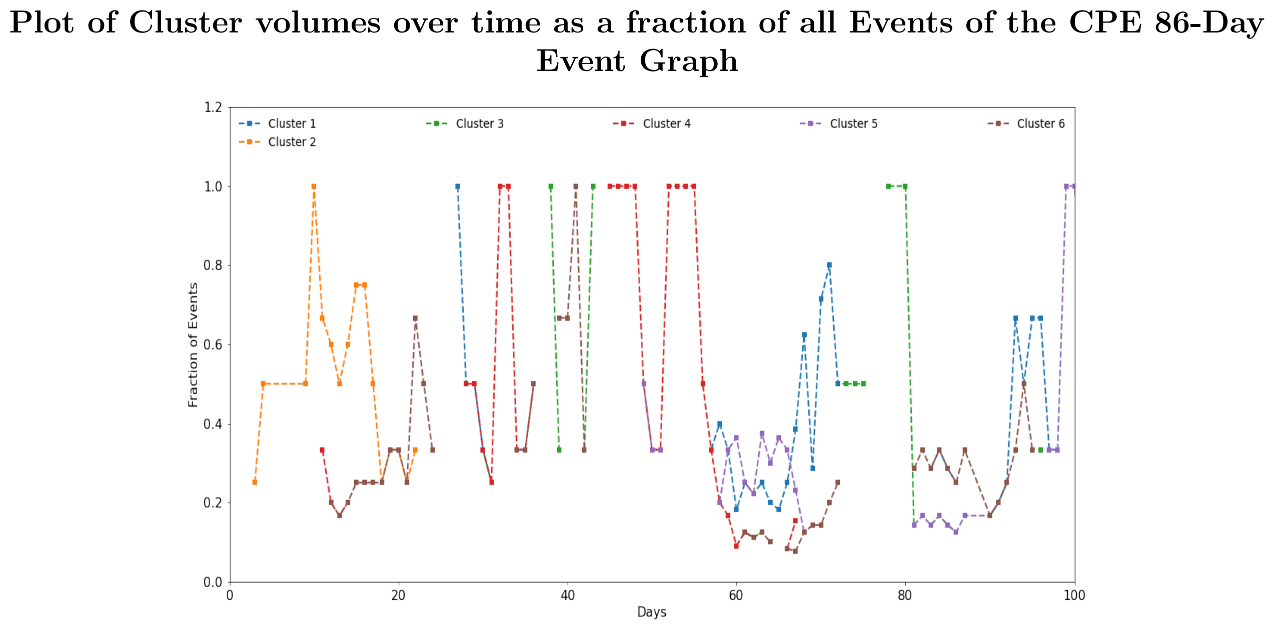


Figure 4.10: Plot of Cluster volumes over time as a fraction of all Events of the CPE 86-Day Event Graph. Certain periods of time are dominated by single cluster's (i.e. Cluster 4 makes up entire network over the period [45,48]. Other times are shared between clusters (consider the time period [60,70]

Overall we are seeing large variability in cluster volumes as a fraction of all events. We have clusters which make up 100% of the entire network at any given point only then to drop to 20% (consider day 80). There also exists periods of time where we can see the inter-play between multiple clusters. Between [60, 70] days for example, we see multiple clusters each making up between 20-30% of the entire network. these periods in time represent when the different types of contact sequences (clusters) are happening at the same points in times. This would suggest 'busy' periods in the Trust, such as around the turn of the new year. We also see periods of time, where there exists no events at all. For example at Day 25. Which could occur for any reason that causes no contact events to occur on a particular day, such as the isolation of all patients on a particular ward (which would prevent 'ABAB').

COVID-19: Cluster Evolution's Over Time

In Figure 4.11 we plot the size of a cluster as a fraction of all events by days (for the period [0,100]).

Plot of Cluster volumes over time as a fraction of all Events of the COVID-19 24-Day Event Graph

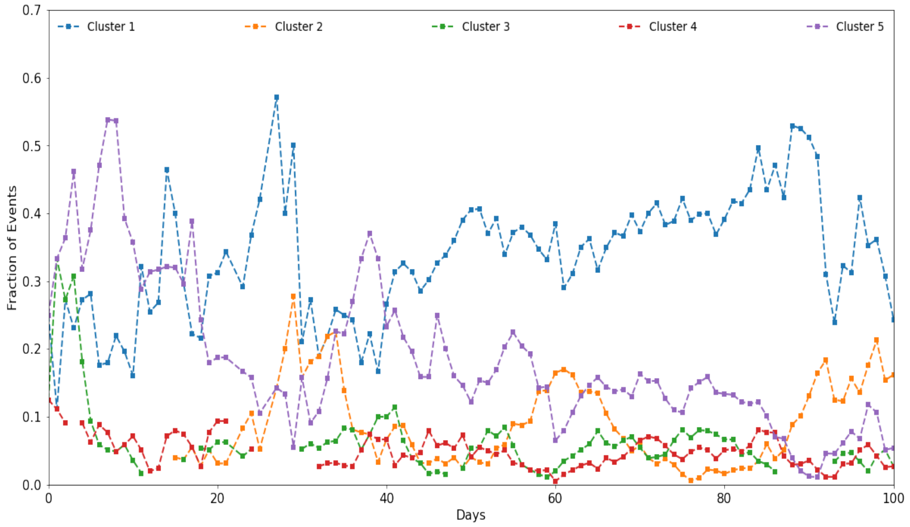


Figure 4.11: Plot of Cluster volumes over time as a fraction of all Events of the COVID-19 24-Day Event Graph. Cluster 1 is consistently dominant throughout the duration, where as cluster's 3 and 4 consistently make up the smallest fraction of events.

Clusters 3 and 4 make up the smallest fraction of the network being relatively constant across time and making up less than 10% at a given point for the majority of the duration (cluster 3 was larger than 10% for the first five days). If the reader recalls, cluster 3 and 4 represented reciprocal events of each other, cluster 3 containing 'ABAC' (Sink Event) whereas cluster 4 contained 'ABCB' (Superspreader Event), therefore this consistency between both clusters over time is no suprise.

Cluster 1 which contains contact sequences which potentially can facilitate pathogen spread very well (widest motif distribution) is consistently very dominant throughout time, in particular in the second half of the duration. This is a concern for the trust as we are seeing contact sequences which facilitate COVID transmission very well, making up the majority of the network for a prolonged period.

Interestingly, cluster 5 which represents repeated interactions across time initially starts off very dominant, but as time progresses decreases and eventually at day 100 makes up the lowest proportion of the network. A possible explanation is that the contact sequences are 'dying' out as time progresses, as a result unfortunately of patients dying, and hence the proportion of events that fall into this cluster reduces over time. Since the only contact structure is 'ABAB' the virus is restricted to just these patients, with no possibility of spreading beyond these patients. Hence if patients are dying, the number of contact sequences are reducing. On the hand, we could be seeing instead that the number of events at a given time in cluster 1 which includes contact sequences that facilitate the spread of diseases very effectively, are increasing as the virus is potentially diversifying over different wards and patients. Thus more events are happening within this cluster as time grows.

4.7 Comparison with time shuffled data

Clustering analysis of the contact sequences (components) of the Δ-t Event Graph has highlighted the diverse behaviour and varying levels of ability for a disease to spread over these contact sequences, but how do we know that these clustered contact sequences are not just based on noise in our data?

To investigate this, we follow *Mellor* [33] and consider an ensemble of two hundred time-shuffled samples of the data. Since we are without a suitable reference model we instead must benchmark our findings using these ensembles of randomised time-shuffled data. To shuffle the data we permute the event times randomly while keeping the node pairs constant. This preserves network structure and the numbers of events between all pairs of nodes but destroys all temporal correlation between events [13].

For each sample we first create an 'average' feature vector x^* for the dataset by considering the entire network as a single contact sequence. Then for each contact sequence (component) feature vector x_c we subtract off the average feature vector

$$x_c^{'} = x_c - x^*$$

This preserves Euclidean similarities between contact sequences and furthermore, the standard deviation σ of $|x^{'}|$ provides a measure of the diversity of outbreaks within the network.

4.7.1 CPE vs Ensemble of Randomised Networks

For the ensemble of randomised CPE networks, the mean and standard deviation are 0.9063 and 0.1254 respectively. In contrast, the mean of $|x|$ for the original CPE data is 1.4591. Comparing this to the ensemble distribution equates to a z-score of 4.4069, see Appendix A.6.

This means that the contact sequences in our CPE data are significantly more diverse than those in a randomised model. Hence we can conclude that the component features we are measuring are significant and not noise in our data, as they are more diverse than we would expect to see at random.

4.7.2 COVID-19 vs Ensemble of Randomised Networks

For the ensemble of randomised COVID-19 networks, the mean and standard deviation are 0.9027 and 0.1260 respectively. In contrast, the mean of $|x|$ for the original data is 1.4042. Comparing this to the ensemble distribution equates to a z-score of 3.9800, Appendix A.6.

Just as with CPE, this means that the contact sequences in our COVID-19 data are significantly more diverse than those in a randomised model. Hence we can conclude that the component features we are measuring are significant and not noise in our data, as they are more diverse than we would expect to see at random.

Since the component features we are measuring are significant and not noise in our data we can infer that, because we are seeing different behaviours across contact events and over time for the Event Graphs, that the decomposition of the temporal network as an Event graph is justified and produces better results than considering the temporal network as a single entity, where we would lose this diversity.

4.8 Final Results and conclusions

In this chapter we successfully clustered the contact sequences of our event graphs based on similarities in the features we described in Chapter 3. We found the optimal number of clusters to be six for the CPE 86-Day Event Graph and five for the COVID-19 24-Day Event Graph. Our aim was to gain an insight into how the underlying contact sequences of these clusters allowed for disease transmission. We looked particularly into the motif distribution (the diversity of the contact structures within a contact sequence) which we found, via the Principle components, to be the distinguishing feature. Upon further analysis, we were able to distinguish distinct categories of sets of clusters based on the likelihood for disease transmission over the contact sequences within each of the clusters. Furthermore, we found that these categories were applicable to both Event Graphs, with both displaying similar sets of clusters. This was a very interesting discovery and one which we based our recommendations to the Trust on, encouraging the Trust to concentrate resources on lessening the duration and diversity of contact sequences within clusters which fell into the 'Highly Effective Transmission Potential' (HETP) Category. We suggested different approaches to tackling this including flagging particular contact events, based on their structure, which encourage the potential for diverse contact sequences and thus the further dispersion of the disease as risky and therefore that the necessary precautions be taken. We also suggested that particular wards, which we defined as hot-spot wards, increase the frequency of testing in an attempt to pick up when a patient is positive quicker than the current methods employed by the Trust. These hot-spot wards, we defined based on the frequency of contact structures to pass through them, i.e. these wards act like hubs in which many event take place.

Finally, we considered how these clusters evolved over time as a fraction of all events. This produced some worrying results, suggesting that in both CPE and COVID-19 clusters which fell into the HETP Category dominated the duration of the event graphs (i.e. that for throughout the whole duration, clusters which fell into HETP made up the majority of events at a given time). This reinforced the need to drastically reduce the diversity and duration of contact sequences grouped in clusters that fell into the HETP category.

Chapter 5

Conclusion

5.1 Summary of Thesis Achievements

In this thesis, we attempt to understand and analyse the contact structures between two completely separate groups of patients who tested positive for their respective diseases. To do this we use the notion of a second-order time unfolded graphical model known as an Event Graph which maps from a given temporal network to the equivalent dual network, where the temporal events of the original network are now the vertices of the Event Graph and edges in the Event Graph are connected to other edges if they have a node in common.

We were generously provided research rights by Imperial College London NHS Trust for two completely separate datasets, composed of the temporal events (Source and Target) and location of those events for the pathogen Carbapenemase Producing Enterobacteriaceae (CPE) and, separately, the virus SARS-CoV-2 (COVID-19) within the Imperial College London NHS Trust. The focus of this study was to model these temporal events between patients as contact sequence (components) of an Event Graph with the aim to gain insight into possible disease transmission over these contact sequences and provide the Trust with recommendations on how to reduce and prevent further transmission.

The decomposition of an Event Graph requires a choice of Δ-t parameter, which thresholds the time between to consecutive events to be connected (Δ-t Adjacency). We introduced our own unsupervised learning method to accurately tune the value of the Δ-t parameter by examining how variations in Δ-t effect the temporal topological structure of the Δ-t Event Graph, by considering four separate temporal and topological measures of the event graph. Our results aligned confidently with the clinical research in both cases.

We then used *Mellor el al.* [33] computational algorithm to build the Event Graph based on our choice of the parameter Δ-t. In order to gain an insight into these contact sequences (components of our Event Graph), we considered twelve temporal and topological features of the Event Graphs and applied hierarchical clustering via Wards method to cluster these contact sequences into groups based on the similarities in the features of each contact sequence.

We found the clustering fell into three distinct categories based upon how likely the contact sequences within a cluster are to potentially spread the disease throughout the Trust. We defined these three categories as 'Highly Effective Transmission Potential' (HETP), 'Mild Transmission Potential' (MTP) and 'No Transmission Potential' (NTP). Clusters which fell into the HETP Category presented contact sequences which were very diverse and have a large number of events, patients involved and locations across the Trust. Intuitively, these represented contact sequences which would facilitate the transmission of a disease very effectively and as such we concluded that these clusters presented a major risk factor to the Trust. We therefore recommended that the Trust focused its resources on reducing the duration and diversity of contact sequences that fall into this category. We suggested that the Trust increases the frequency of testing, for the respective diseases, on what we defined as 'Hot-Spot' wards, which are the wards the contact sequences, within a set of clusters, are most frequently involved within. We also suggested that certain contact structures such as $ABAC$ (Superspreader Event) and $ABCB$ (Sink Event), which inflate the diversity of a contact sequence be flagged by the Trust as potentially risky and only undertaken with special precautions if necessary. We believe that implementing these considerations within the Trust will be an effective way of reducing the diversity and duration of patient contact sequences in general but most importantly those that could potentially facilitate the transmission of disease most effectively.

The MTP Category, which group's cluster's composed of contact sequences which could potentially facilitate the transmission of a disease throughout the Trust but are far more contained compared to those in the HETP Category. We found that on average the duration of these contact sequences were far less than those in the HETP Category, which suggested that the Trust was doing relatively well at picking up when patients in these contact sequences are positive for there respective disease, isolating and ending the contact sequence relatively quickly. Of course, contact sequences within these clusters still facilitate the possibility of disease transmission throughout the Trust and therefore should be of concern too. Contact sequences within these clusters, have a tighter motif distribution (i.e. contain less diverse contact structures) with motif types 'ABCB' (Sink Event) and 'ABAC' (Superspreader Event) being the second majority after 'ABAB' (Repetitive Event). Similarly to the above recommendations, we suggested that the Trust does everything it can to prevent these two types of contact structures from being allowed. These contact structures occur when patients are transferred between wards or a new admission is allowed on the ward. We suggest the Trust not only flags these events as risky for disease transmission but also includes precautionary considerations if the event is necessary. For example, if a patient needs to be transferred between wards or a patient needs to be admitted to a new ward, they could under go a pre-isolation period, before being admitted to new ward, where extra testing is undertaken to ensure the patient is not infectious.

The final category, which we defined as having no transmission potential contained clusters of contact sequences which were composed of only repetitive interactions ('ABAB') between two patients on the same ward. These contact sequences, can support disease transmission but because of the composition, the same two patient interacting on the same ward, there is no risk of further disease spread beyond these two patients because it is contained. However these patients should not be discarded and must ensure that testing is continued. Note that, clusters in this category contained the lowest average duration with respect to all over clusters, which suggests the trust is managing when a patient is testing positive in these clusters and isolating them, ending the contact sequence, very well. With all things considered, clusters in this category pose the least threat to the trust overall and we recommend the Trust continue with its current methods for testing patients who are involved in these repetitive contact sequences.

5.2 Model Limitations

Possible limitations to our model are:

- The assumption of only dyadic interactions. (Discussed in Section 5.3)
- Choosing to hierarchically cluster via Wards method.

Our choice to hierarchically cluster via Wards method was perhaps not given enough thought and in hindsight using a Gaussian Mixture model instead would be better. On reconsidering the data in the PCA plots, see Appendix A.4 and Appendix A.5 we see that in fact many of the clusters are ellipsoidal. *Everitt* [12] showed that Ward's method often leads to misclassifications when the clusters are distinctly ellipsoidal rather than spherical. Therefore, we may have a small number of missclassifications within our clusters, however we still believe that the general results of this project are justified given the analysis conducted.

5.3 Future Work

A possible avenue for future work would be the extension to a Hyper-Event Graph. In this thesis, as in many studies of networks, we considered the assumption of pairwise interaction of nodes (dyadic interactions). However in some cases this assumption is an oversimplification. Consider a digital communication network such as text messaging. A text message can be sent to numerous individuals in a single event, this is a non-dyadic interaction event. In the context of our thesis, a non-dyadic event would be when a single patient interacts with multiple other patients at the same point in time. This is a realistic scenario, considering a ward has multiple patients. Formally a Hyper-Event Graph is defined as:

Definition 5.1 ***Hyper-Event Graph.***

Let $V \subseteq \mathbb{N}$ *be a set of nodes,* $T \subseteq \mathbb{R}_0^+$ *a set of times,* $D \subseteq \mathbb{R}_0^+$ *be a set of duration's, and* E *a set of temporal hyper-events. A temporal hyper-event graph is defined by the quadruple* $G = (V, T, D, E)$ *where temporal hyper-events take the form*

$$e_i = (U_i, t_i, \delta_i)$$

where $U_i \subseteq V$*. For* directed *temporal hyper-event graphs hyper-events take the form*

$$e_i = (U_i, V_i, t_i, \delta_i)$$

where $U_i, V_i \subseteq V$ *and* $U_i \in V_i = \emptyset$ *[35].*

To extend our current model to a Hyper-Event Graph we would continue with the assumption that a node can only be involved in at most one event at any given time but now assume non-dyadic events are possible and therefore that a node (patient) may have multiple edges during the event. This is a realistic assumption. Consider a patient on a ward with multiple other patients, that patient may interact with multiple patients on that ward in a given event, however that one patient cannot interact in multiple events i.e. cannot be on two different wards at the same time.

Appendix A

Appendix

A.1 Algorithm for Event Graph Decomposition

Below we outline the steps to construct the event graph and perform a temporal decomposition by thresholding the edge weights. This algorithm is implemented in the eventgraphs Python package, available freely online [36]. Beginning with an empty graph:

1. For each node x construct a time-ordered sequence of events $S_x = (e_k)_{k=1}^{|S_x|}$ such that $e_i \in S_x$ if and only if $x \in \{u_i, v_i\}$, i.e. the set of events for which x is a participant.

2. For each consecutive pair of events e_k, e_{k+1} in S_x, add an edge from e_k to e_{k+1} in the event graph with weight $t_{k+1} - t_k$, where t represents time of the event. Repeat this process for each node in the temporal network.

The edges of the event graph can then be thresholded to removed all edges over a set value Δ-t. The weakly-connected components of the event graph can then be found through standard methods by considering the corresponding undirected network.

A.2 Scores Indicating Optimal Number of Clusters.

Here we include the plots of the Silhouette, Davies Boulidn and Calinski Harabasz Scores used to indicate the optimal number of clusters for the components of each Event Graph.

Scores indicating the optimal number of clusters for CPE 86-Day Event Graph

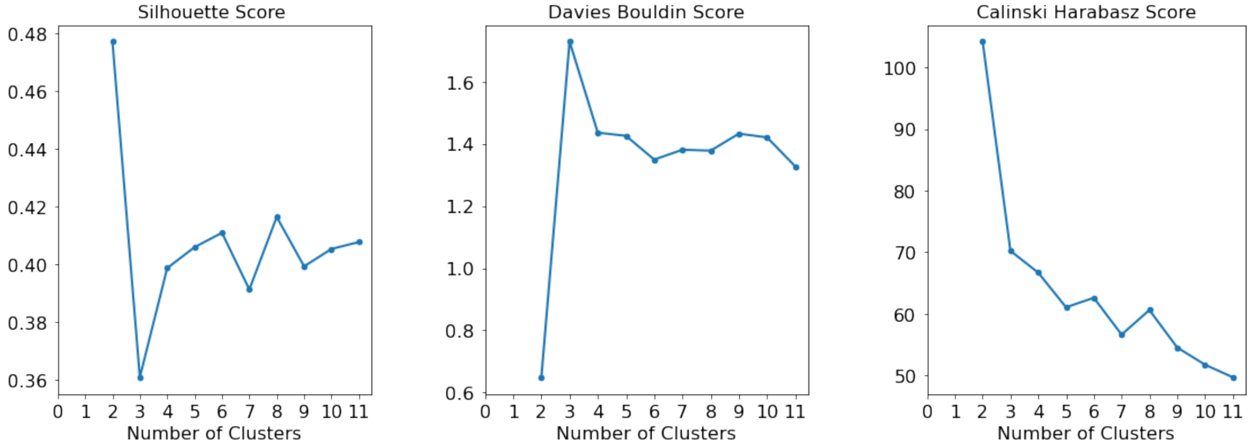


Figure A.1: Silhouette, Davies Boulidn and Calinski Harabasz Scores used to indicate optimal number of clusters for the components of the CPE 86-t Event Graph

Scores indicating the optimal number of clusters for COVID-19 24-Day Event Graph

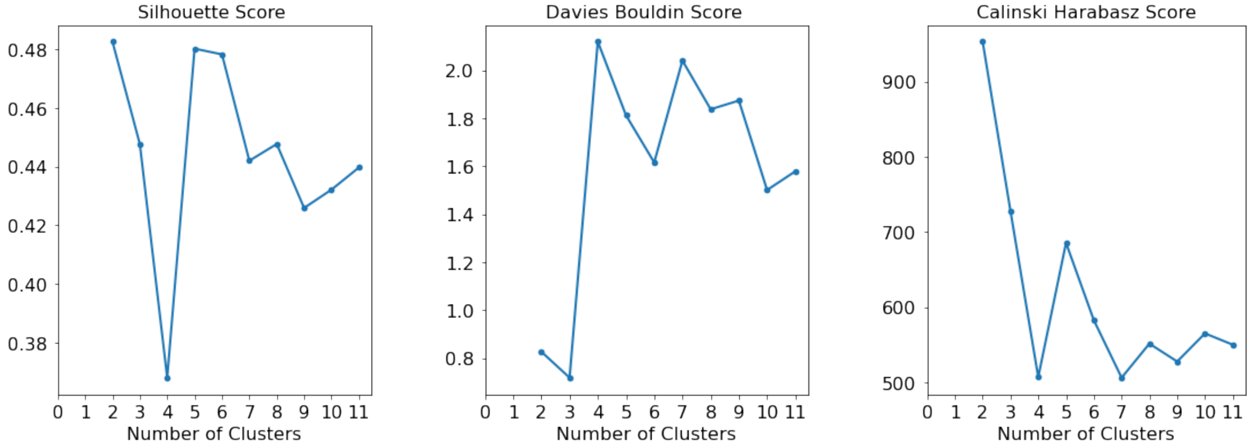


Figure A.2: Silhouette, Davies Bouldin and Calinski Harabasz Scores used to indicate optimal number of clusters for the components of the COVID 24-t Event Graph

A.3 Plots of the Cluster Volume Evolution's of the CPE 86-Day Event Graph.

Below are the plots of cluster volumes over time as a fraction of all Events of the CPE 86-Day Event Graph for the whole duration, 659 days. Split into interval periods of 100 days.

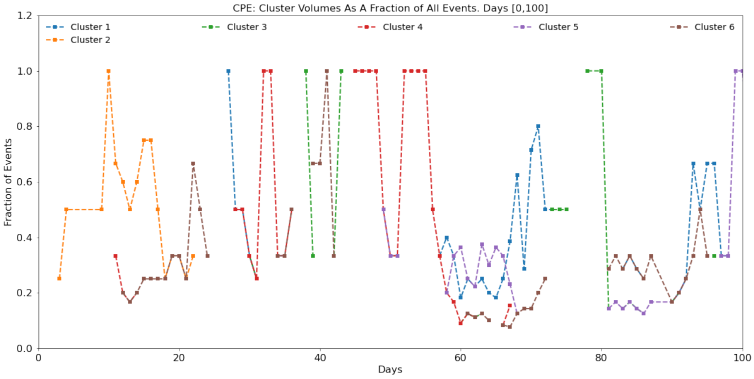


Figure A.3: Plot of cluster volumes over time as a fraction of all Events of the CPE 86-Day Event Graph for the period [0,100] Days.

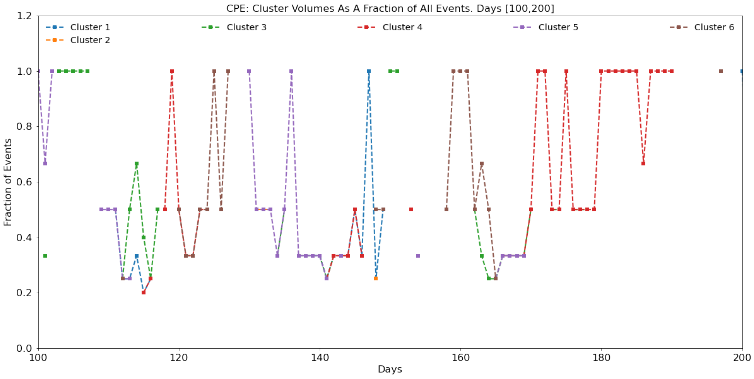


Figure A.4: Plot of cluster volumes over time as a fraction of all Events of the CPE 86-Day Event Graph for the period [100,200] Days.

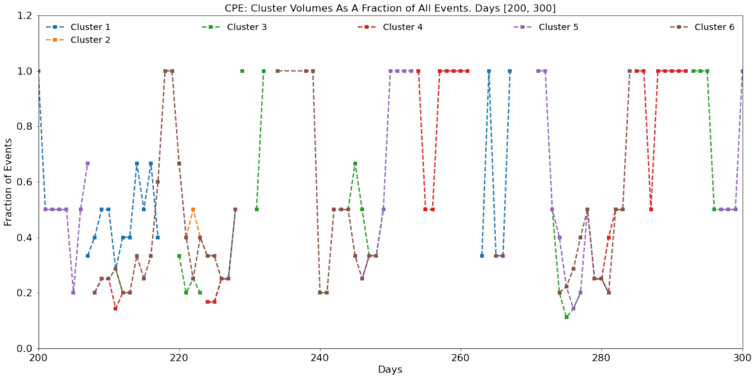


Figure A.5: Plot of cluster volumes over time as a fraction of all Events of the CPE 86-Day Event Graph for the period [200, 300] Days.

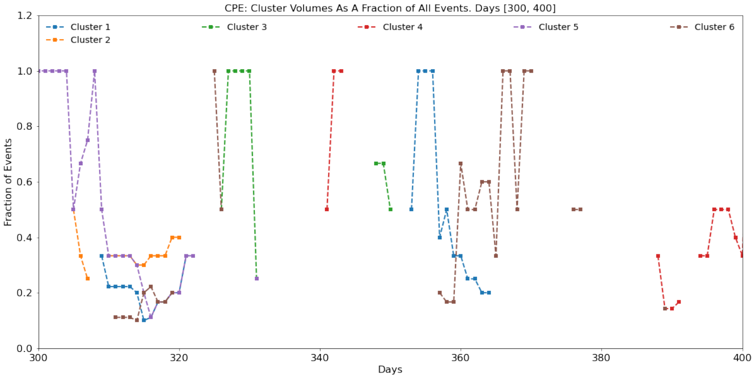


Figure A.6: Plot of cluster volumes over time as a fraction of all Events of the CPE 86-Day Event Graph for the period [300, 400] Days.

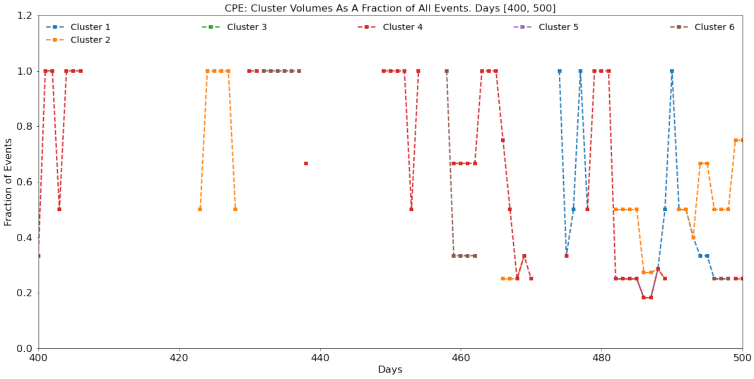


Figure A.7: Plot of cluster volumes over time as a fraction of all Events of the CPE 86-Day Event Graph for the period [400, 500] Days.

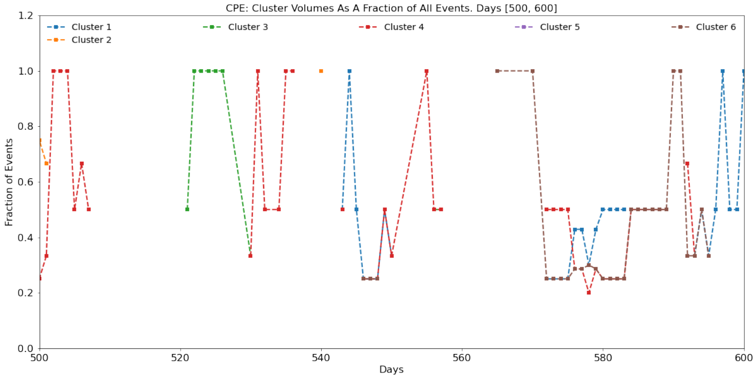


Figure A.8: Plot of cluster volumes over time as a fraction of all Events of the CPE 86-Day Event Graph for the period [500, 600] Days.

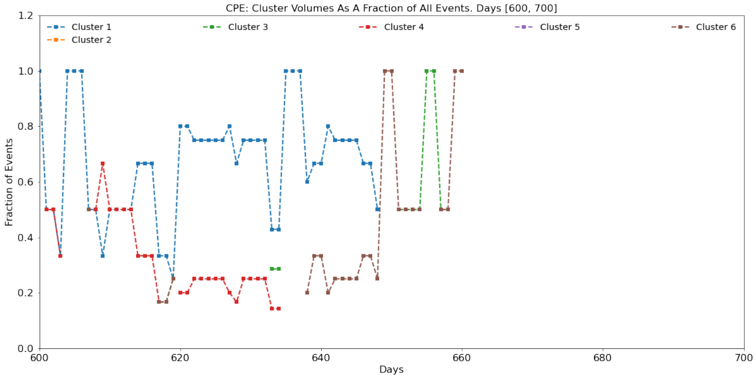


Figure A.9: Plot of cluster volumes over time as a fraction of all Events of the CPE 86-Day Event Graph for the period [600, 700] Days. Duration ends at day 659

A.4 CPE: PCA Plots in all Three Dimensions

Below are the scatterplots of the contact sequences in the first three components of the PCA reduced feature space for the CPE 86-Day Event Graph.

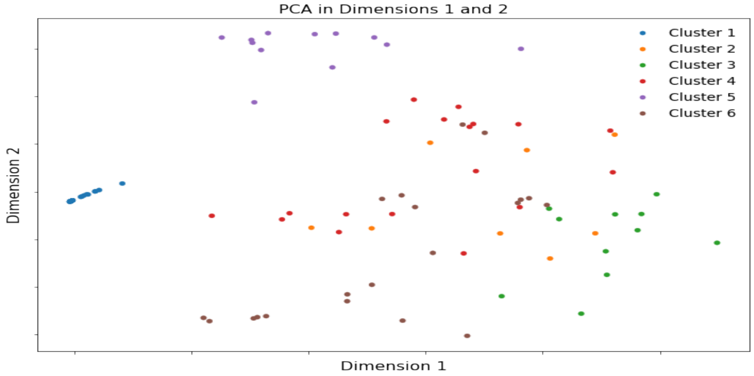


Figure A.10: Scatterplot of the contact sequences in the first and second components of the PCA reduced feature space for the CPE 86-Day Event Graph

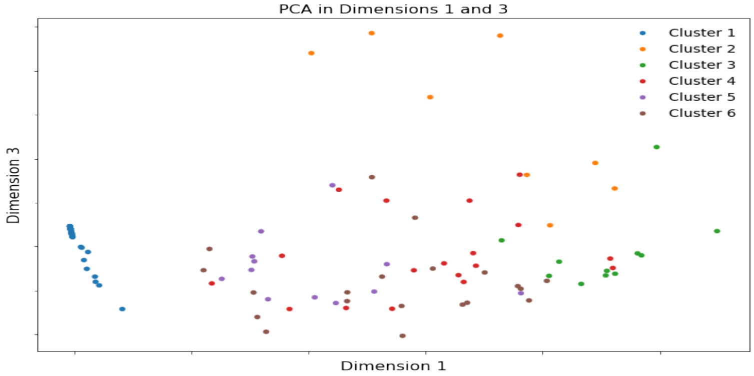


Figure A.11: Scatterplot of the contact sequences in the first and third components of the PCA reduced feature space for the CPE 86-Day Event Graph

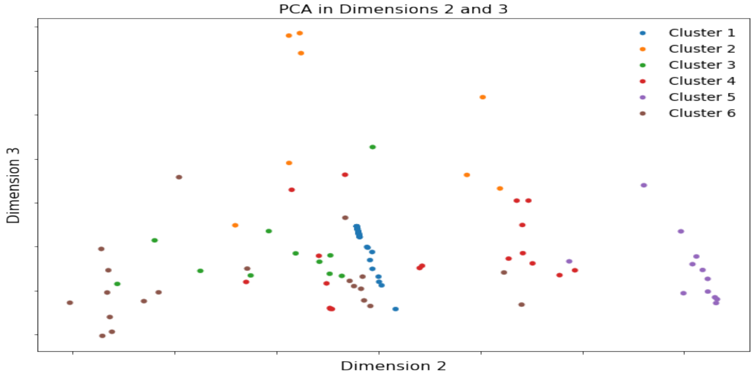


Figure A.12: Scatterplot of the contact sequences in the second and third components of the PCA reduced feature space for the CPE 86-Day Event Graph

A.5 COVID-19: PCA Plots in all Three Dimensions

Below are the scatterplots of the contact sequences in the first three components of the PCA reduced feature space for the COVID-19 24-Day Event Graph.

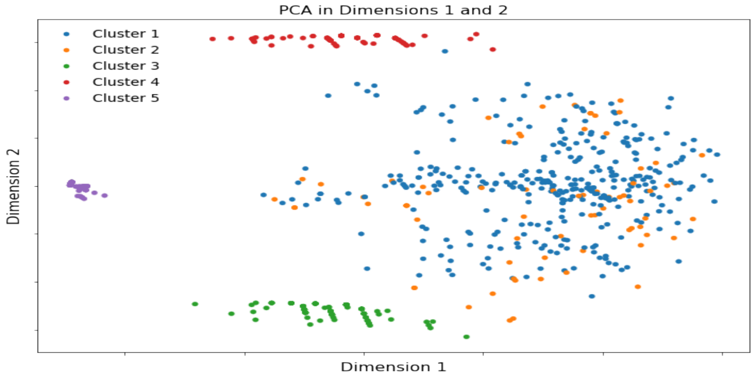


Figure A.13: Scatterplot of the contact sequences in the first and second components of the PCA reduced feature space for the COVID-19 24-Day Event Graph

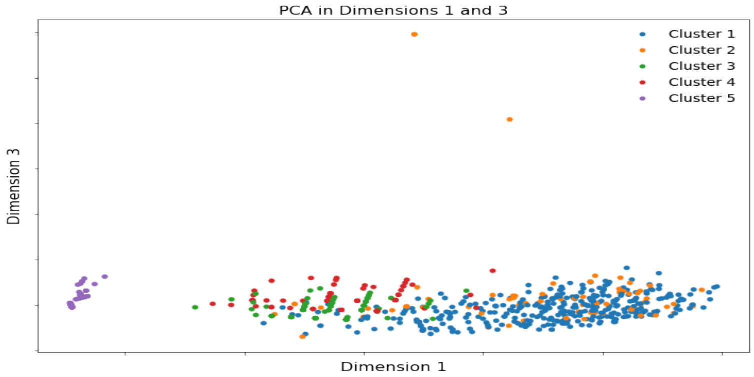


Figure A.14: Scatterplot of the contact sequences in the first and third components of the PCA reduced feature space for the COVID-19 24-Day Event Graph

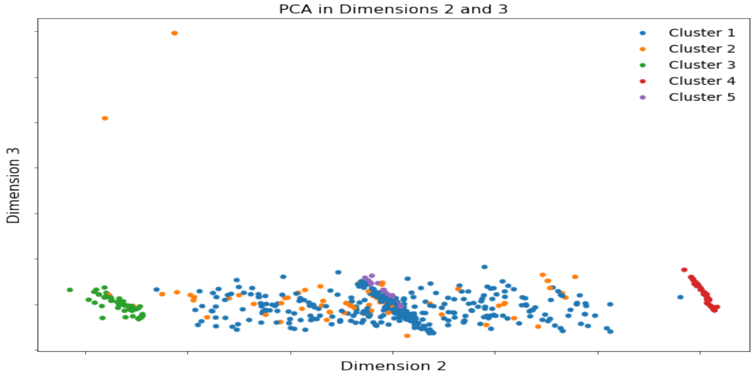


Figure A.15: Scatterplot of the contact sequences in the second and third components of the PCA reduced feature space for the COVID-19 24-Day Event Graph

A.6 Normal Dist Plots

Below are the Probability density distributions for the ensembles of randomised networks for both CPE (Figure A.16) and COVID-19 (Figure A.17). The red cross indicated the mean value of the original datasets respectively. Both provided large z-scores and as such we concluded that in both cases the component features we are measuring are significant and not noise in our data, since they are more diverse than what we would expect to see at random.

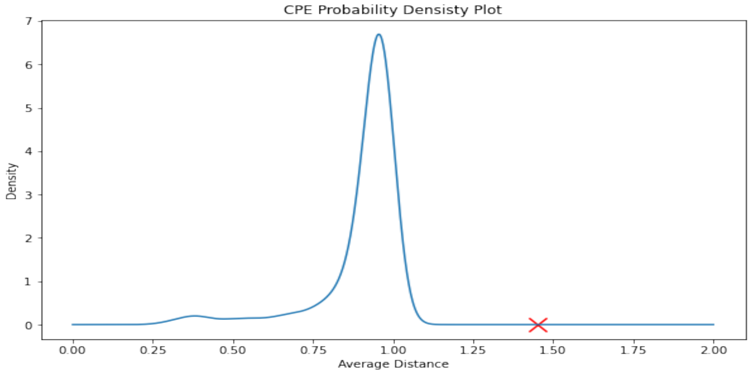


Figure A.16: Probability density distribution for the ensemble of randomised CPE networks. Red cross indicates the mean value of the original CPE data. Z-score is significant.

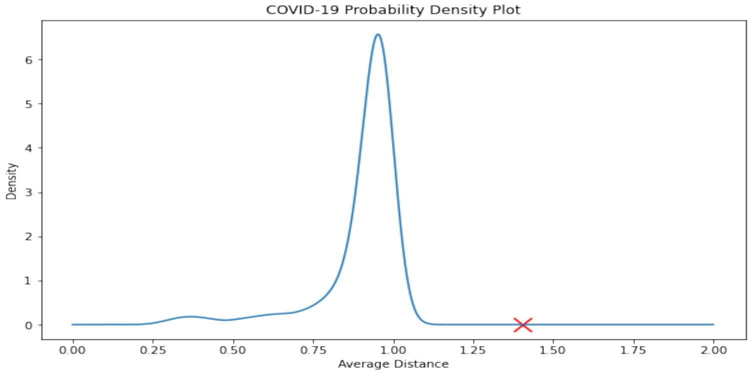


Figure A.17: Probability density distribution for the ensemble of randomised CPE networks. Red cross indicates the mean value of the original CPE data. Z-score is significant.

Bibliography

[1] Dillon C Adam, Peng Wu, Jessica Y Wong, Eric HY Lau, Tim K Tsang, Simon Cauchemez, Gabriel M Leung, and Benjamin J Cowling. Clustering and superspreading potential of sars-cov-2 infections in hong kong. *Nature Medicine*, 26(11):1714–1719, 2020.

[2] NTJ Bailey. General epidemics. *The mathematical theory of infectious diseases and its applications*, pages 81–133, 1975.

[3] Béla Bollobás. *Modern graph theory*, volume 184. Springer Science & Business Media, 2013.

[4] Robert A Bonomo, Eileen M Burd, John Conly, Brandi M Limbago, Laurent Poirel, Julie A Segre, and Lars F Westblade. Carbapenemase-producing organisms: a global scourge. *Clinical infectious diseases*, 66(8):1290–1297, 2018.

[5] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[6] Thomas F Coleman and Jorge J Moré. Estimation of sparse jacobian matrices and graph coloring blems. *SIAM journal on Numerical Analysis*, 20(1):187–209, 1983.

[7] Jie Cui, Fang Li, and Zheng-Li Shi. Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*, 17(3):181–192, 2019.

[8] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[9] Nicolaas Govert De Bruijn. A combinatorial problem. In *Proc. Koninklijke Nederlandse Academie van Wetenschappen*, volume 49, pages 758–764, 1946.

[10] Austin Derrow-Pinion, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire, et al. Eta prediction with graph neural networks in google maps. *arXiv preprint arXiv:2108.11482*, 2021.

[11] Ken TD Eames and Matt J Keeling. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proceedings of the national academy of sciences*, 99(20):13330–13335, 2002.

[12] Brian S Everitt. Unresolved problems in cluster analysis. *Biometrics*, pages 169–181, 1979.

[13] Laetitia Gauvin, Mathieu Génois, Márton Karsai, Mikko Kivelä, Taro Takaguchi, Eugenio Valdano, and Christian L Vestergaard. Randomized reference models for temporal networks. *arXiv preprint arXiv:1806.04032*, 2018.

[14] Michael C Grant, Luke Geoghegan, Marc Arbyn, Zakaria Mohammed, Luke McGuinness, Emily L Clarke, and Ryckie G Wade. The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (sars-cov-2; covid-19): a systematic review and meta-analysis of 148 studies from 9 countries. *PloS one*, 15(6):e0234765, 2020.

[15] Wei-jie Guan, Zheng-yi Ni, Yu Hu, Wen-hua Liang, Chun-quan Ou, Jian-xing He, Lei Liu, Hong Shan, Chun-liang Lei, David SC Hui, et al. Clinical characteristics of coronavirus disease 2019 in china. *New England journal of medicine*, 382(18):1708–1720, 2020.

[16] Ki-ichiro Hashimoto. Zeta functions of finite graphs and representations of p-adic groups. In *Automorphic forms and geometry of arithmetic varieties*, pages 211–280. Elsevier, 1989.

[17] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[18] Petter Holme. Network reachability of real-world contact sequences. *Physical Review E*, 71(4):046119, 2005.

[19] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[20] Alison H Holmes, Luke SP Moore, Arnfinn Sundsfjord, Martin Steinbakk, Sadie Regmi, Abhilasha Karkey, Philippe J Guerin, and Laura JV Piddock. Understanding the mechanisms and drivers of antimicrobial resistance. *The Lancet*, 387(10014):176–187, 2016.

[21] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395(10223):497–506, 2020.

[22] Christopher JR Illingworth, William L Hamilton, Ben Warne, Matthew Routledge, Ashley Popay, Chris Jackson, Tom Fieldman, Luke W Meredith, Charlotte J Houldcroft, Myra Hosmillo, et al. Superspreaders drive the largest outbreaks of hospital onset covid-19 infections. *Elife*, 10:e67308, 2021.

[23] Ian Jolliffe. Principal component analysis. *Encyclopedia of statistics in behavioral science*, 2005.

[24] Matthew J Keeling, Mark Newman, Albert-László Barabási, and Duncan J Watts. The effects of local spatial structure on epidemiological invasions. In *The Structure and Dynamics of Networks*, pages 480–488. Princeton University Press, 2011.

[25] David Kempe, Jon Kleinberg, and Amit Kumar. Connectivity and inference problems for temporal networks. *Journal of Computer and System Sciences*, 64(4):820–842, 2002.

[26] Mikko Kivelä, Jordan Cambe, Jari Saramäki, and Márton Karsai. Mapping temporal-network percolation to weighted, static event graphs. *Scientific reports*, 8(1):1–9, 2018.

[27] Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, and Jari Saramäki. Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(11):P11005, 2011.

[28] Renaud Lambiotte, Vsevolod Salnikov, and Martin Rosvall. Effect of memory on the dynamics of random walks on networks. *Journal of Complex Networks*, 3(2):177–188, 2015.

[29] Renaud Lambiotte, Lionel Tabourier, and Jean-Charles Delvenne. Burstiness and spreading on temporal networks. *The European Physical Journal B*, 86(7):1–4, 2013.

[30] Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582, 2020.

[31] UK Government Legislation. Coronavirus legislation uk government. https://www.legislation.gov.uk/coronavirus, 2021.

[32] Georgios Meletis. Carbapenem resistance: overview of the problem and future perspectives. *Therapeutic advances in infectious disease*, 3(1):15–21, 2016.

[33] Andrew Mellor. Analysing collective behaviour in temporal networks using event graphs and temporal motifs. *arXiv preprint arXiv:1801.10527*, 2018.

[34] Andrew Mellor. The temporal event graph. *Journal of Complex Networks*, 6(4):639–659, 2018.

[35] Andrew Mellor. Event graphs: Advances and applications of second-order time-unfolded temporal network models. *Advances in Complex Systems*, 22(03):1950006, 2019.

[36] Dr. A Mellor. 'eventgraphs' python package. https://github.com/empiricalstateofmind/eventgraphs, 2018.

[37] Lauren Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society*, 44(1):63–86, 2007.

[38] Lauren Ancel Meyers, MEJ Newman, Michael Martin, and Stephanie Schrag. Applying network theory to epidemics: control measures for mycoplasma pneumoniae outbreaks. *Emerging infectious diseases*, 9(2):204, 2003.

[39] Yin Mo, Anastasia Hernandez-Koutoucheva, Patrick Musicha, Denis Bertrand, David Lye, Oon Tek Ng, Shannon N Fenlon, Swaine L Chen, Moi Lin Ling, Wen Ying Tang, et al. Duration of carbapenemase-producing enterobacteriaceae carriage in hospital patients. *Emerging infectious diseases*, 26(9):2182, 2020.

[40] Mark Newman. *Networks*. Oxford university press, 2018.

[41] Patrice Nordmann, Laurent Dortet, and Laurent Poirel. Carbapenem resistance in enterobacteriaceae: here is the storm! *Trends in molecular medicine*, 18(5):263–272, 2012.

[42] Ashwin Paranjape, Austin R Benson, and Jure Leskovec. Motifs in temporal networks. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 601–610, 2017.

[43] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.

[44] René Pfitzner, Ingo Scholtes, Antonios Garas, Claudio J Tessone, and Frank Schweitzer. Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks. *Physical review letters*, 110(19):198701, 2013.

[45] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[46] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011.

[47] Ingo Scholtes, Nicolas Wider, and Antonios Garas. Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities. *The European Physical Journal B*, 89(3):1–15, 2016.

[48] Ingo Scholtes, Nicolas Wider, René Pfitzner, Antonios Garas, Claudio J Tessone, and Frank Schweitzer. Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. *Nature communications*, 5(1):1–9, 2014.

[49] Muhammad Adnan Shereen, Suliman Khan, Abeer Kazmi, Nadia Bashir, and Rabeea Siddique. Covid-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of advanced research*, 24:91, 2020.

[50] Juliette Stehlé, Alain Barrat, and Ginestra Bianconi. Dynamical and bursty interactions in social networks. *Physical review E*, 81(3):035101, 2010.

[51] E. Tacconelli and N.Magrini. Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. 2021.

[52] Maddalena Torricelli, Márton Karsai, and Laetitia Gauvin. weg2vec: Event embedding for temporal networks. *Scientific reports*, 10(1):1–11, 2020.

[53] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[54] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’networks. *nature*, 393(6684):440–442, 1998.

[55] Felix Wong and James J Collins. Evidence that coronavirus superspreading is fat-tailed. *Proceedings of the National Academy of Sciences*, 117(47):29416–29418, 2020.

[56] Worldometers.info. Global coronvirus statistics. https://www.worldometers.info/coronavirus/, 2021.

[57] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, et al. A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269, 2020.

[58] Renyi Zhang, Yixin Li, Annie L Zhang, Yuan Wang, and Mario J Molina. Identifying airborne transmission as the dominant route for the spread of covid-19. *Proceedings of the National Academy of Sciences*, 117(26):14857–14863, 2020.

[59] Yi-Qing Zhang, Xiang Li, Jian Xu, and Athanasios V Vasilakos. Human interactive patterns in temporal networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(2):214–222, 2014.

[60] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature*, 579(7798):270–273, 2020.

[61] Lirong Zou, Feng Ruan, Mingxing Huang, Lijun Liang, Huitao Huang, Zhongsi Hong, Jianxiang Yu, Min Kang, Yingchao Song, Jinyu Xia, et al. Sars-cov-2 viral load in upper respiratory specimens of infected patients. *New England Journal of Medicine*, 382(12):1177–1179, 2020.