Dynamics and Organisation in Complex Systems: From Cytokines to Cities



Sofia Medina

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2025

To Doctor José Carlos Medina, who taught me to always follow what inspired me.

Para mi querida Laura, que me cuida siempre y se sentía orgullosa de mí desde mi primer aliento.

Para Yeyé, en nuestros corazones siempre estábmos cerca. Y en cada mate que tomo, ahí estamos, nieta y abuela juntas. Será como decíamos, ya cuando nos encontremos seré la Doctora Medina.

Y para vos, que siempre me decís qué linda vida que tengo, y yo te digo, "será por tanto que rezás por mi". Para Elisa

Acknowlegements

First and foremost, I would like to thank my wonderful supervisors Eamonn Gaffney and Renaud Lambiotte. Without your support, encouragement, and seemingly endless knowledge, this thesis would not have been possible. I feel incredibly grateful to have had such patient mentors. Your guidance is truly what allowed me to grow as a mathematician and a professional. You always provided me with honest and gentle advice, helping me navigate the endless challenges of the DPhil. Somehow, there was always room in every meeting for a few jokes and some laughs together.

Another thanks are reserved for Timothy LaRock and Nicola Pedreschi, who I am fortunate to have encountered as my colleagues, but more importantly, I am lucky to count among my friends. Your insightful feedback and willingness to entertain countless questions have been invaluable in this journey. The most important lesson that you taught me is that first and foremost, mathematicians are people. Curious, kind, and wonderfully human people.

Thank you to David Egea, an endless reservoir of knowledge and a listening ear, who shared curiosity for science and for life and brought cheekiness to every discussion, making this DPhil so much richer.

I would also like to thank my mentors along my journey that brought me to Oxford, Richard Bertram and Santiago Schnell. They believed in me, guided me, and encouraged me to pursue my dreams. I could never express in words how immensely grateful I am for the opportunities that you have provided me. For the ways that you have changed my life and confidence and curiosity that you instilled in me, I will forever be deeply thankful.

I am incredibly grateful for my time at the Linacre Ladies that Lift. Thank you to Catherine and Jenni for their endless support and encouragement. I also would like to thank Shez, Carrie, Isabel, Karen,and Taylah who Lifted me up and reminded me how strong I am. With every laugh and smile shared, you made the whole world as light as a feather.

I would like to express my thanks to the Wolfson Centre for Mathematical Biology and the Networks group. My wonderful friends who accompanied me in this wild experience Rohit, Bridget, Jesus, Chloe, Christoph, Kara, Alex, Josh, and Giulia. Also to my officemates Arianna, Lara, Yuan, and Solveig. Especially to that one who I love to pretend is my nemesis, Vedang. You all were my company in this journey. You made me smile in moments when I stumbled or doubted myself. Its because of the support and the friendship that you gave me with such ease that while I grew as a mathematician, I grew as a person. No thank you to Steven Aaardvark.

I would like to thank Lewis, Paddy, Ali and Billie who made my first house feel truly like a home and filled it with lifelong friend to love. Cristiana and Zeev who gave me their tender friendship and showed up for me in my most difficult moments.

I would like to thank Thales, my connection through the years for reminded me of who I was, who I am, and who I am becoming. Thank you to my wonderful companions Montse, Ava, and Rayyan for sharing with me countless joys and walking beside me in my moments here. I would like to thank Valentina and Shazia. Thank you for being the shoulders I rested my head on, the friends I confided in, and my wisest council. Your friendship is truly a beacon of light.

Rajwa, thank you for coloring my Oxford journey in warm and vibrant tones. You anchored me with your kindness and your support, and you brightened all my days with your mischief. This truly wouldn't have been the same without you and I am so grateful for you.

To my friends from home, Jennica, Vanessa, Ethan, Agustina, Anastasia, Giovanna, Brian, Kirsty, Jacob, and Louis. You are and always will be a part of me. You remind me that with true friendship, the world is small. Thank you for your support and for the connections that we keep, tend to, and treasure. Saying that I count you as my people are the true great achievements of my life.

Gracias a Tomás, que es una flecha que apunta a un mejor futuro. Infinitas gracias a mi familia, que me acompaña en todos mis logros y siempre camina conmigo.

Finally, to my parents. Thank you for always believing in me, for always supporting me, and for being the people I can always turn to. Your examples are my guiding light. Thank you for showing me the true meanings of resilience, perseverance, and most importantly of love.

Abstract

Complex systems, with their intricate web of interacting components, are ubiquitous across diverse domains. In this thesis, we employ mathematical models and develop novel methodologies to study such systems in a variety of applications. The work is organized into three parts, each addressing systems at progressively larger scales.

In the first part, we examine a network of immune system signalling molecules extracted from *in vitro* gut biopsy data and assess the dynamical influence of individual components on each other. We further extend the model to include *in vivo* system features. We identify system regulators and note key differences in the system models upon the introduction of *in vivo* features.

In the second part of this thesis, we analyse trends in mobile phone application traffic following major events. We detect spatio-temporal changes in application traffic and characterise trends in application usage. Our findings reveal at fine-grained scale the impact of application features and emphasise city-specific user application preferences. We develop a methodology for spatial tracking of traffic response through a null model, demonstrate the emergence of a radial data traffic application pattern following an event shock, and track its subsequent relaxation.

In the third part, we develop a novel methodology to analyse connectivity and reachability in systems modelled by directed hypergraphs, in order to account for multi-body interactions. Building on the notion of reactive cascades, we apply this method to chemical reaction data, demonstrating the structure of the data and giving insights into chemical organisation.

Taken together, this thesis contributes new methods for the study of complex systems, revealing structural patterns and their effects within datasets, and introducing methodological tools and system-level insights to support further investigation.

Contents

1 Introduction **1**
1.1 Introduction to Complex Systems 1
1.2 Thesis aims 4
1.3 Statement of Originality 6
1.4 General Preliminaries 7
1.4.1 Kullback-Leibler Divergence 8
1.4.2 Clustering Analysis 9
1.4.2.1 The k-means algorithm 9
1.4.2.2 k Selection 10
1.4.3 Maximum Likelihood Estimation 11
1.4.4 Akaike Information Criterion 12
1.4.5 Sensitivity analysis 12
1.4.6 Networks 13
1.4.7 Latin Hypercube Sampling 15
1.4.8 Biological Experimental Approaches 17

I The Extraction and Analysis of Cytokine Networks and Associated Dynamics in Inflammatory Bowel Disease **19**

2 Overview and Preliminaries **20**
2.1 Overview 20
2.2 State of the Art and Model Assumptions 22
2.2.1 Model Assumptions and Experimental Details 23
2.2.1.1 Units of concentration 24
2.2.2 Network extraction technique 24

3 The Cytokine Network for IL-23 Regulation **27**
3.1 Introduction . . . 27
3.2 Methods . . . 30
3.2.1 Quantitative summary of model . . . 30
3.2.2 Robustness to Noise . . . 32
3.2.3 The in vitro and in vivo systems . . . 33
3.2.4 Sensitivity of IL-23 . . . 34
3.2.5 Model Reduction . . . 34
3.2.6 IL-1β, IL-10 and IFNγ . . . 35
3.3 Results . . . 35
3.3.1 Robustness to Noise . . . 35
3.3.2 The Steady State . . . 38
3.3.3 Sensitivity analysis . . . 41
3.3.4 Model Reduction via removal of IL-1α . . . 47
3.3.5 The effect of IL-1β . . . 49
3.3.6 IL-10 and IFNγ . . . 52
3.4 Discussion . . . 53
3.5 Conclusions and Future Directions . . . 57

II Temporal and Spatial Organisation of Mobile Traffic Response to Catastrophe **59**

4 Overview **60**
4.1 Overview . . . 60

5 Detection of Anomalous Spatio-temporal Patterns of App Traffic in Response to Catastrophic Events **64**
5.1 Introduction . . . 64
5.2 Methods . . . 67
5.2.1 Data . . . 67
5.2.2 Detecting Spikes in App Traffic Volume . . . 69
5.2.3 Temporal Clustering . . . 70
5.2.4 Spatial Analysis . . . 71
5.2.4.1 Traffic Spikes in Space . . . 72
5.2.4.2 Quantifying Spatial Effects . . . 73
5.3 Results . . . 75

5.3.1 Temporal Analysis: Anomalous Application Traffic in Response to the Fire . . . 75
5.3.2 Two-Day Analysis: Short and Long Term Behavioural Clusters and Patterns . . . 77
5.3.2.1 Patterns on the Day of Fire . . . 77
5.3.2.2 Patterns on the Day after Fire . . . 83
5.3.3 Cluster Selection and Feature Values . . . 86
5.3.3.1 Features, day of fire . . . 86
5.3.3.2 Features, day after fire . . . 90
5.3.4 Joint Two-Day clusters . . . 93
5.3.5 Spatial Analysis of Application Traffic Response . . . 95
5.3.5.1 Twitter Traffic in Paris . . . 96
5.3.5.2 Lyon, a Further Study of Spatial Spreading Patterns 102
5.4 Discussion . . . 107
5.5 Conclusions and Future Directions . . . 110

III Directed Hypergraph Exploration and the Chemical Space 112

6 Overview and Preliminaries 113
6.1 Overview . . . 113
6.2 Preliminaries . . . 115
6.2.1 Hypergraphs and Directed Hypergraphs . . . 115
6.2.2 Molecular and Synthetic Complexity . . . 116
6.2.2.1 Assembly Theory . . . 117

7 Cascade Processes in Directed Hypergraphs and the Architecture of the Chemical Space 119
7.1 Introduction . . . 119
7.2 Reactive Cascade and Reactive Walks . . . 120
7.2.1 Properties of the Reactive Cascade . . . 124
7.3 Application . . . 130
7.3.1 Dataset and Data Processing . . . 131
7.3.2 Directed Hypergraph Null Model Randomisation . . . 133
7.3.3 Results . . . 133
7.3.3.1 Hypergraph Statistics . . . 133

7.3.3.2 Edge Metrics 136
7.3.3.3 Node Metrics 141
7.3.4 Node Selection Reactive Cascade 145
7.4 Discussion 146
7.5 Conclusion and Further Directions 148

8 Conclusion **150**
A Appendix A 153
A.1 Constraints on network extraction due to the multi-start approach 153
A.2 Parameter values for regulatory cytokines 153
B Appendix B: Part II 155
Appendix B: Part II Appendix 155
B.1 Full List of Applications Available in NetMob2023 Dataset . . 155

Bibliography **156**

Chapter 1

Introduction

1.1 Introduction to Complex Systems

Across diverse fields and disciplines, complex, interacting systems frequently arise. The field pertaining to the study and approaches to understanding such systems and their fundamentally complicated nature is known as *complex systems* or *complexity science*.

While there is no universally accepted definition of a system that is complex, such systems are characterised by the interconnection of their often heterogeneous components and the specifics of the dependencies on their parts. Complex systems also often experience feedback, whereby their behaviour adjusts at different time-points according to the conditions on which they interact, as well as their own state. Nonlinearity and stochasticity also contribute to system complexity, with complex systems evolving over time in ways that can be difficult to predict [64, 117, 162, 219].

Complexity itself can be defined across scales of size or time, with the possibilities of complex layers developed by the specifics of the system itself. Interactions of components at small scale may impact the organisation and dynamics of the system at a larger scale. The interactions of these small parts often lead to a phenomenon fundamental to the study of complex systems known as *emergence* [159, 206, 64, 162, 222]. That is, system behaviours that emerge from the interactions of a complex system in the sense that they cannot be fully explained by the properties of individual parts themselves *a priori*. One such example of emergence is pattern formation. For example, mathematical models of cellular motion through reaction diffusion systems can be used to simulate animal coat patterns such as zebra stripping and giraffe spotting [155], with a general and broad interest in modelling focused on the self-organisation of interacting species [139, 99]. Cooperative behaviours among individuals also lead to the emergence of behaviour phenomena such as bird flocking [44] and fish swarming

[1]. Emergence can also be framed in more abstract concepts, for example: how waves in a medium lead to our perception of sound or how our experience of consciousness emerges from the arrangement of electrical signals in the brain [159, 206].

We need not look far to understand the ubiquity of complex systems. We can interrogate physical systems such as planetary motion or the synchronisation of fireflies through the lens of complex systems [193, 183]. We can interpret systems through complexity in the understanding of interactions and behaviours found in social networks or the connections of the internet [172, 125]. In an attempt to gain insight into phenomena, generally explore, and understand interacting systems, data is generated from observations of such systems. In turn, mathematical models are developed to parse this information and approximate the real system behaviours in order to study complex systems. Attempts at understanding complex systems have been developed through a variety of sub-fields that aim to explain emergence and system behaviour through specific approaches. These methods are often designed to interrogate and extrapolate results from model systems to understand emergent behaviours and general trends.

Complexity and mathematical modelling go hand in hand [162]. Within topics like biology and medicine, the mathematical models applied to study complex systems are crucial to developing understanding [155, 106]. For example, infectious diseases and epidemics can be simulated through mathematical models. Epidemic models can track population interactions with dynamical equations that describe interactions between compartments, such as Susceptible-Infected-Recovered models (SIR) [15, 155, 132], or use data regarding physical contact patterns to theoretically investigate and predict the impact of epidemic interventions [65]. The breadth of application of mathematical modelling ranges across biological disciplines, from the development of the Hodgkin-Huxley model to describe the electrophysiology of neurons [92] to mathematical models giving insight into the complex mechanisms of the retina and ear [106, 121]. Further mathematical modelling in biology is extensive and has been applied to organs such as the heart [167, 203, 166] and the brain [217, 127, 115], fundamental structures such as tendons [69, 137, 43] and bones [178, 177, 192], and biological mechanisms such as oxygen transport [151, 229].

Mathematical models also probe the complexity of chemistry [230, 234]. The Law of Mass Action describes chemical reaction rates allowing for the characterisation of chemical equilibria [226, 155]. Building upon this, enzymatic reactions can be represented by the Michealis-Menten reaction mechanism, which provides a quantitative framework for catalytic processes, with extensive work undertaken to describe

enzyme catalysis under the assumption of a quasi-steady state [197, 195, 60]. Chemical reactions can be modelled stochastically at low numbers of individual molecules, tracking fluctuations and oscillating chemical trajectories [77, 62, 198, 225]. Further at larger numbers, chemical equations can be modelled by mean-field continuum models through differential equations to describe reaction-diffusion processes [208, 31].

Another subfield through which complexity is studied is known as *mobility* or *human mobility* which seeks to understand phenomena around human motion, such as human migration, urban development, air transportation, and traffic forecasting, among many other applications [81, 17]. Studies use data from censuses, GPS locations, and even surveys [17]. Their applications and investigation range across scales. For example, individual-level actions may be modelled using Levy flights, where an individual's displacement length is taken from a power law distribution resulting in mostly small displacements, with rarer large steps [188, 35]. Individual actions are also modelled by the exploration and preferential return (EPR) model, where individuals return to previously visited locations with some probability and explore new locations with the complementary probability. These probabilities can evolve over time, taking into account visitation frequency and recency [211, 18].

At larger scales, population-level actions may be modelled as flows according to models based on Newton's law of gravity, specifically emphasising the influence of distance and population size (analogous to mass) on mobility [207]. Further yet, models may incorporate cross-scale effects, such as the d-EPR model, which extends the EPR framework by taking into account collective influence on exploration such that individuals are more likely to preferentially explore population dense locations [174]. Through the use of models of human mobility, a variety of emergent phenomena can be captured, including segregation [194] and gentrification [147].

Techniques within complex systems are as wide ranging as their applications. Agent-based models use individual 'agents', which may have assigned heterogeneous characteristics and decision-making abilities. These agents and their characteristics represent the behaviour of individuals or small system components such as cells, demonstrating the emergence of behaviours and trends arising from their interactions [194, 61, 147]. The use of dynamical and nonlinear systems are another ubiquitous technique in the description of the interactions of complex systems, capturing behaviours from interactions between components to the time evolution of species [197, 155, 106, 132, 159]. Network theory also provides a rich source of methodology for understanding complex systems, where the interdependencies of elements can be modelled as links between them [222].

Complexity science lies at the forefront of many fields and techniques, driving innovation and probing the complicated features of nature itself. The study of complex systems seeks to unveil and analyse the dynamics of interconnected structures across scales. The tools and applications derived within the umbrella of complexity science capture system percolation and collective behavioural trends. Challenges to studying complex systems are inherent to their nature. For example, system models should be sufficiently detailed to accurately describe the system and capture its emergent phenomena, while avoiding over-parametrisation and retaining the ability to interpret the model with insight. The multi-scale nature of complex systems also makes them difficult to capture well and adds further layers of computational complexity. Further, complex systems may be highly dependent on initial conditions and inherently stochastic, creating further challenges. Finally, complexity science often requires high resolution data, which can be difficult to attain. Investigators in the field therefore attack problems from multiple angles, seeking to uncover interactions and describe the dynamics and organisation of these interacting systems with a breadth of tools.

1.2 Thesis aims

In this thesis, we will look at three systems through the lens of complex systems. Our goal is to develop and test methods that unveil structures of interacting systems and demonstrate how data organisation impacts system dynamics. We aim to reveal data patterns across various real-world systems through our analyses and explore data propagation on these systems. These aims will be developed in three parts, regarding studies of ascending information size. We begin with small-scale cytokine interaction networks, from which we will gain insight into how each component leads to larger dynamical behaviour and unveil the impact of structures that are initially unclear. We will then present a mobility study on mobile phone application data patterns, showcasing trends in human behaviour emerging from responses to catastrophic events. Finally, we will culminate our journey through complex systems in an analysis of the organisation of large-scale chemical reaction datasets. This thesis will be presented as follows.

In **Part I: The Extraction and Analysis of Cytokine Networks and Associated Dynamics in Inflammatory Bowel Disease**, we present a study that opens the goal of this thesis of describing dynamics and data organisation of interacting systems. We begin in the small scale through the cytokine interaction network,

noting that only sparse experimental data exists. This is an immunological network that signals pro- and anti-inflammatory responses. This system is of interest to the scientific and mathematical modelling community in the understanding and treatment of auto-immune diseases, as well as their simulation. In Chapter 2, we give an introduction to the study of cytokine networks and the network extraction model from which our research study is derived. In Chapter 3, we probe the organisation of the network topology and study its dynamics by exploring a specific application of the extraction of a cytokine interaction network with limited data. We thoroughly scrutinise the dynamics of the system and introduce *in vivo* features to the system, observing how closely model results align with *in vitro* predictions after the introduction of additional assumptions. This study thoroughly explores the regulation of the promising therapeutic cytokine target IL-23 and emphasises the importance of caution in the translation of *in vitro* model predictions to *in vivo* models. We also explore the robustness of the network extraction methodology, assessing the recovery of the structure of the cytokine interaction network as presented by previous studies. This assessment plays a critical role in applying network extraction techniques to a biological system with sparse data.

Part II: Temporal and Spatial Organisation of Mobile Traffic Response to Catastrophe, presents an exploration of spatio-temporal patterns of mobile phone application traffic due to catastrophic events. In Chapter 4, we motivate and provide background on mobility studies, discussing their various applications and the insights they provide to diverse groups of interest. In Chapter 5, we present how information propagates in space and time, paying particular attention to trends of return-to-baseline activity following a perturbation in application traffic. We also demonstrate patterns of similar behaviour across platforms, noting short and long term behaviours of applications and city-specific user preferences. Finally, observing spatial patterns of data spread with respect to the epicentre of catastrophe, we construct a null model of spatial data organisation in time for comparison with observed data. This study further develops the aims of uncovering patterns in dynamical systems and revealing trends in data organisation, describing emergent collective human behaviour in response to catastrophe.

In **Part III: Directed Hypergraph Exploration and the Chemical Space**, we move to the question of selecting appropriate models for complex systems and developing insightful methods for those models. Our objective is to enrich the network

framework used to encode and analyse multi-body interactions. For this purpose, we develop a walk-based methodology on directed hypergraphs. This methodology probes the connectivity of directed hypergraphs as well as the influence of hyperedges and nodes. Further, we will explore the chemical space through its representation as a directed hypergraph of chemical reactions. We apply our methodology and developed metrics to describe features of the chemical space as well as individual reactions and molecules. We reveal the interacting structure of the reaction system through our analysis. The properties of the chemical space and large scale data-driven hypergraphs are of interests to both the scientific community and the network science community [131, 187]. In Chapter 6, we will develop the background and description of tools used in describing directed hypergraphs and the investigation of the chemical space. We then introduce our novel method and detail its application in Chapter 7. This study further reiterates the objective of revealing the structure of an interacting system and developing methodologies through which to interrogate complex systems.

Finally, we conclude in Chapter 8 by providing an overview of the contributions described in this thesis.

1.3 Statement of Originality

Portions of the DPhil research undertaken were collaborative. In the following points, I specify which chapters involved collaborations and to what extent.

- **Chapter 3: The Cytokine Network for IL-23 Regulation**. Eamonn Gaffney helped conceptualise and supervise this project. Mark Coles and Holm Uhlig provided insight into the immunological and scientific background of this work. The network extraction methodology utilised and evaluated in this section was developed in [98] by corresponding authors. Experimental data in relation to this project were obtained by [12], with the initial cytokine network extracted by [98]. All further analysis, code development, and manuscript writing was done by me. Chapter 3 was sponsored by GSK plc and in collaboration with GSK research scientists. A manuscript based on Chapter 3 has been submitted to the Journal of the Royal Society Interface and is pending review.

- **Chapter 5: Detection of Anomalous Spatio-temporal Patterns of App Traffic in Response to Catastrophic Events**. This chapter was a joint collaboration with Nicola Pedreschi and, in minor proportion, with Shazia'Ayn

Babul, Timothy LaRock, and Rohit Sahasrabuddhe. This work, of which I am the first author, has been published in the journal EPJ Data Science and is publicly available at [148]. All three latter authors contributed to the editing and some writing of the manuscript, with Shazia'Ayn also contributing to the initial conceptualisation and Rohit writing a function for data extraction. The bulk of the work was undertaken by Nicola and I, with Nicola in a supervisory role and I the driving author. Nicola played an equal role in the conceptualisation of the null model. I wrote the code for the detection of traffic spikes and feature extraction. I ran this code on data from the city of Paris and Nicola ran it on the remaining cities, due to the distribution of the available data across our computers. K-means clustering was implemented by Nicola using code available from a previous project. He and I both contributed to the development of code for spatial measurement of traffic spikes. While he and I both helped write the manuscript, I wrote the majority of it. I also wrote the code for the creation of the null model, KL-divergence over time, and the adaptation of spatial measurements in one direction within city limits. Renaud Lambiotte was a senior supervisor on this project.

- **Chapter 7: Cascade Processes in Directed Hypergraphs and the Architecture of the Chemical Space**. Timothy LaRock and Renaud Lambiotte helped conceptualise this project. David Egar provided insight into chemical knowledge and scientific background of this work. Timothy LaRock wrote the function that downloaded the data from the ORD. After the original functions were written by me, Timothy also provided support and suggestions to improve code efficiency for the hypergraph extraction and cleaning process, as well as for the functions tracking levels in the reactive cascade. All further analysis, code development, and manuscript writing was done by me. A manuscript pertaining to this work is in preparation.

1.4 General Preliminaries

In the following sections, we introduce the main concepts from information theory, data science, networks, and biology that will help us develop methods to analyse our complex systems throughout the thesis. We note that more particulars regarding methodologies and states of the art will be further developed within other chapters.

1.4.1 Kullback-Leibler Divergence

The Kullback-Leibler Divergence, also known as KL divergence and sometimes referred to as relative entropy, is a statistical distance measure that measures how much a probability distribution Q differs from a probability distribution P [116, 90, 49].

The Kullback-Leibler divergence, in its discrete form, is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in X} p(x) \log \left(\frac{p(x)}{q(x)} \right), \tag{1.1}$$

where P and Q are discrete probability distributions over a random variable $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, and $p(x)$ and $q(x)$ denote their respective probability mass functions.

KL divergence is derived from the information-theoretic concept of information content, or "surprisal" [199]. This describes the information gained by an event E. When P(E) is close to 1, its surprisal is low and little information is gained by the occurrence of an event E. When P(E) is close to 0, surprisal is high, with more information given by its occurrence. This is described uniquely by the log function

$$I(E) = \log \left(\frac{1}{P(e)} \right), \tag{1.2}$$

or

$$I(E) = -\log(P(e)). \tag{1.3}$$

We can then define the concept of entropy, specifically Shannon entropy, which quantifies the information content needed to describe outcomes drawn from the probability distribution of a random variable $\mathcal{X}$ as

$$H(X) \coloneqq -\sum_{x \in X} p(x) \log(p(x)), \tag{1.4}$$

which arises from the expected value of the information content of $\mathcal{X}$ [199].

Following these first-principles, we suppose we have events from a probability distribution P and use a model distribution Q to encode events from P. The information content from the use of Q is therefore

$$H(P, Q) = -\sum_{x \in X} p(x) \log(q(x)), \tag{1.5}$$

which is the average surprise given by observing P with its corresponding outcome probabilities $p(x)$ using the model distribution of Q with the consideration of Q's surprisal of its outcomes within the log function. This is known as cross-entropy [199, 49].

Hence, when using the model Q instead of the true model P, the difference between the cross entropy, H(P,Q), and the entropy of P itself, H(P), is the KL divergence. This gives the additional cost value of the use of Q instead of P, a measure of the difference between the two probability distributions. That is

$$D_{\mathrm{KL}}(P \parallel Q) = H(P,Q) - H(P) = \sum_{x \in X} p(x) \log \left(\frac{p(x)}{q(x)} \right). \tag{1.6}$$

We note that KL divergence is asymmetric, non-negative, unbounded above and zero if and only if $P = Q$ [116].

1.4.2 Clustering Analysis

In data analysis, clustering refers to the grouping or classification of data into disjoint subsets or clusters. Clustering allows for the partition of data into groups of similar characteristics [56, 66, 83, 138].

1.4.2.1 The k-means algorithm

The clustering method that will be used is the k-means algorithm, a standard method of data clustering, which partitions n data points into K clusters $C_1, \ldots, C_k$, $k \in \{1, 2, \ldots, K\}$. We note that this clustering satisfies $C_k \subseteq \{1, \ldots, n\}$, $C_k \cap C_{k'} = \emptyset$ $\forall k \neq k'$, and $\bigcup_{k=1}^{K} C_k = \{1, ..., n\}$ [56]. The k-means algorithm is an unsupervised, iterative clustering algorithm. Each of the k clusters is represented by the mean value of objects in the cluster, called the centroid of the cluster. Initially, each of the k centroids is randomly initialised and each data point is assigned to the nearest centroid cluster based on a Euclidean distance metric. The cluster centroid is then updated by computing the mean distance of the cluster centroid for all data points within the current cluster, after which the centroids are reassigned to the within-cluster average. The objective of k-means clustering is to minimise the distance between within-cluster data points and the corresponding centroid, through successive iteration until no change is made to the centroid assignment, or a predetermined number of iterations is completed [56, 241, 138, 103]. A visualisation of the k-means algorithm after initialisation and convergence is shown in Figure 1.1. We note that k-means selection is potentially sensitive to the random initialisation of cluster centroids, which can be mitigated or addressed through the use of multi-start. The k-means algorithm is also sensitive to the number of clusters k, which must be selected *a priori*. The methodology used in this work for the selection of k, known as the 'elbow method' is detailed in the following section.

1.4.2.2 k Selection

The partition and classification of data using the k-means algorithm is highly dependent on the number of clusters, k, used in the algorithm. For example, a partition where $k = n$ is not particularly informative, however on the other extreme a partition of $k = 2$ that splits the data into two halves may also not be an appropriate analysis of the given data. We therefore aim to locate the number of clusters for a dataset that remain informative without overfitting to the data. In the following work, we use the 'elbow method' to locate an optimum number of clusters for our dataset. The elbow method is a graphical methodology based on the calculation of the sum of square error distance between the centroid and all within-cluster data points for various k values. The elbow methodology iterates through a sequence of k values, typically $k \in 1, ..m$ where $m \in \mathbb{Z}$, performing the k-means algorithm on the data for each value of k. The total within-cluster sum of square error for each value of k is considered the value of the cost or error for using this k value. The error cost is then plotted as a function of the number of clusters. It is expected that the dataset will experience substantial decreases in its error cost function as the number of cluster increases, as well as plateau in the cost function as the addition of further clusters provides minimal changes to within cluster distances. These curves are sometimes referred to as *inertia curves*. The "optimum" number of clusters is then determined by the 'elbow' of the plotted curve, which corresponds to the point of its maximum curvature before a plateau in value [51, 26, 214, 87]. A visualisation of the use of the elbow method for the selection of optimal k in the k-means algorithm is shown in Figure 1.1.

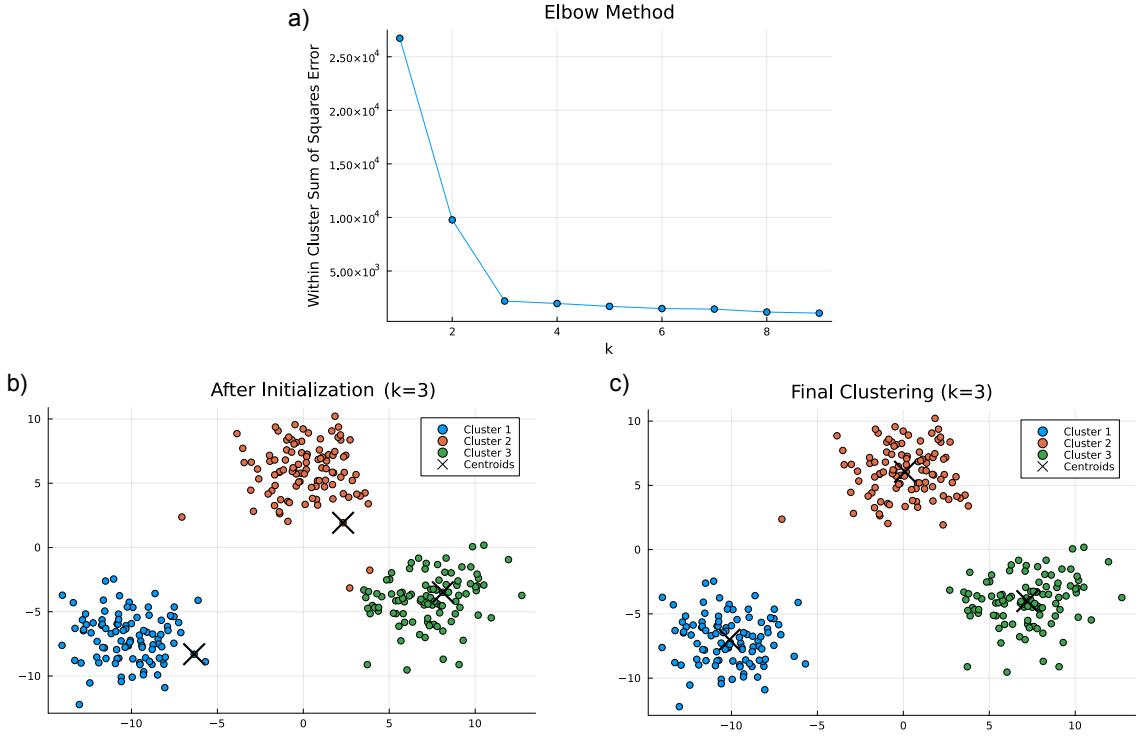

Figure 1.1: 2D example of the k-means algorithm with corresponding elbow method for the selection of k. **a)** Plot of different within cluster sum of squares error cost for the selection of each value of k. Clearly, the optimum k value to be selected is 3. **b)** Centroids and their respective assigned data points after initialisation at $k = 3$. **c)** Centroids and their respective assigned data points after the convergence of the k-means algorithm at $k = 3$.

1.4.3 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a parameter estimation method that maximises the likelihood that chosen parameter values are the most probable for observing given data under a fixed model [56, 97, 157].

We denote $f(x;\theta)$ as a probability distribution of a random variable X which depends on a parameter θ. We can define $X_1, X_2, ..., X_n$ as independent random variables where $\mathbf{x} = (x_1, ..., x_n)$ are observations of the values of the random variables. Assuming independence of observations, the joint probability distribution is therefore

$$f(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{1.7}$$

The likelihood function is defined as this joint probability distribution, but viewed as

a function of θ for fixed $\mathbf{x}$ as,

$$\mathcal{L}(\theta) = f(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{1.8}$$

Commonly, the log-likelihood, $\ell(\theta) = \log(\mathcal{L}(\theta))$ is used for convenience as the log is a monotone transformation and allows for the use of summations as opposed to products. Maximum likelihood estimation is therefore the maximisation of this likelihood, or log-likelihood function, through numerical optimisation algorithms or analytical methods. Equivalently, one may employ the minimisation of the negative log-likelihood function to return the desired parameters value of the model. Often, the minimisation of $-2\log(\mathcal{L}(\theta))$ is used as it is asymptotically equivalent to the χ^2 distribution under the regularity conditions of Wilks' Theorem, making it convenient for statistical inference [97, 233, 52, 57].

1.4.4 Akaike Information Criterion

The Akaike Information Criterion (AIC) is a statistical inference technique used for model selection during inference. The AIC takes into account the fit of the model based on maximum likelihood estimation and penalises based on the number of model parameters needed to fit the model. In this way, the AIC balances model complexity with model fit in order to account for possible over-fitting. The AIC value for model selection is

$$\text{AIC} = 2p - 2\log(\mathcal{L}(\hat{\theta})) = 2p - 2\ell(\hat{\theta}), \tag{1.9}$$

where p is the number of parameters in the model, $\mathcal{L}(\hat{\theta})$ is the likelihood for the given model evaluated at the maximum likelihood estimate $\hat{\theta}$, and $\ell(\hat{\theta})$ is the maximised log-likelihood [4, 41, 98].

1.4.5 Sensitivity analysis

Global sensitivity analysis is implemented using the method eFAST (extended Fourier amplitude sensitivity test), a decomposition method using Fourier analysis that varies input parameters, returning two sensitivity measures for the effect of a parameter on the measurement of a variable of interest [143, 191, 97, 190]. The first order sensitivity measure, S_i, gives sensitivity via the ratio of the variation due to each parameter i individually as a fraction of total variance. The second measure, total-order sensitivity or S_{Ti}, is calculated by first accounting for the variance of all parameters except i,

the complementary set of i. The leftover variance is assumed to be variance due to the non-linear interactions between i and the complimentary set of i. The total-order sensitivity, or S_{Ti}, is then calculated as the first order measure, S_i, plus the sensitivity not accounted for in the complementary set of i. This incorporates the impact of a parameter via its interaction with other parameters into the measure of S_{Ti} [143, 190].

During eFAST, each input parameter is assigned a unique frequency with the values for this parameter defined along a sinusoidal periodic function, or *search curve*. The model is evaluated at a user-defined number of points along the search curve, denoted N_s, which controls the resolution of the evaluation. A Fourier transformation is then applied to the model evaluation output. Since each parameter was assigned a unique frequency, the contribution of each parameter to the variance can be isolated from the Fourier component at its frequency. This procedure is then repeated at a user defined N_r times, denoted the *resampling size*, where each sinusoidal function is treated with a random phase shift in order to reduce bias. The methodology also introduces a dummy parameter that is not associated with the model equations to act as a control variable that measures artifact due to the algorithm, for example due to sampling resolution. The total number of evaluations is therefore N_s x N_r x $p + 1$, where p is the number of parameters and the addition of one corresponds to the dummy parameter [143, 190, 97].

1.4.6 Networks

A network is a system of nodes and edges. Generally, the nodes, or vertices, represent encoded information objects and the edges, or links, between nodes represent the association or relationship between these units [118, 163]. Networks emerge in many fields with examples ranging from social networks, where nodes may be people and edges may be relationships [244], technological networks where nodes may be web-pages and edges hyperlinks to other web-pages [171], and even biological networks where nodes may be genes and edges represent their regulation interactions [104]. Network edges can be undirected, for example in the mutual association 'friendship', or directed, as in the example of the authorship citation relation 'cites'. Edges may be weighted to represent strength of interaction and signed to represent positive and negative interactions such as up-regulation or down-regulation [118]. Edges may also connect vertices to themselves, called self-edges. Networks, also referred to mathematically as *graphs* are defined as,

$$\mathcal{G} = (V, E), \tag{1.10}$$

where the graph $\mathcal{G}$ is a collection of nodes or vertices $V = \{v_1, v_2, v_3, .., v_n\}$ and edges $E = \{e_1, e_2, e_3, ..., e_m\}$. Each edge in this network is defined by a pair of nodes $e = (v_i, v_j) \in E$. We note that these networks can be considered *pairwise networks*, given the pairwise relation between vertices. For undirected networks, the order of v_i and v_j are interchangeable, however for directed networks, the pair (v_i, v_j) indicates a link from vertex v_i to vertex v_j while the pair (v_j, v_i) indicates a link from vertex v_j to vertex v_i. We note that directed networks may contain both the pairs (v_i, v_j) and (v_i, v_j), indicating a reciprocal relation between the vertices v_i and v_j [163, 118, 112].

One property of a node is its *degree*. In an undirected network, the degree of a node is the number of edges connected to it. In a directed network, the degree of a node is conceptualised as its *in-degree*, the number of in-going edges, and its *out-degree*, the number of its out-going edges. For example, in a directed authorship citation network where the vertices are academic papers, the in-degree of the vertex v_a is the number of other papers that cite paper a and the out-degree of the vertex v_a is the number of other papers that paper a cites [118, 163]. Nodes with high degrees relative to other network nodes are known as *hubs* [163].

Networks are a robust way to represent dynamical processes. For example, one can define a dynamical system upon the edges of a network through the use of differential equations, illustrating the interaction between vertices over time according to the equations governing their edge interactions. One may also define a *walk* on a network, which is a sequence of nodes such that a successive node is connected to the previous node via an edge. For the case of directed networks, edges may only be traversed in the direction a node points to, from the out-edge of one node to the in-edge of its connected node. A walk may intersect itself, i.e. revisit a node in the sequence. A walk that does not intersect itself is considered a *path* through the network. Often, one of the simplest cases of a dynamical process on a network is a *random walk*, which begins at a node i and traverses nodes through edges with some probability associated with the traversal of an edge. Walks and paths give insight into the structure or topology of a network, from which one may define metrics such as shortest paths from a start node to an end node, investigate the multiplicity of paths between nodes, and determine overall connectivity of a network [112, 118, 163]. *Connectivity* refers to the existence of paths between a pair of nodes. In the graph $\mathcal{G}$, two nodes v_i and v_j are said to be *connected* if there exists a path from v_i to v_j. The graph $\mathcal{G}$ itself is said to be connected if there exists a path from all vertices to all other vertices. This is straightforward within the undirected network. However given the directionality of edges within a directed network, it may not be possible to reach all network vertices through edge

traversal. Within a directed network, two nodes v_i and v_j are said to be *weakly connected* if there exists a path between the two nodes where the directionality of the network edges are ignored. Two nodes are said to be *strongly connected* if there exists a path from v_i and v_j as well as a path from v_i and v_j. Likewise, a directed graph, or digraph, is said to be weakly connected if there exists a path from all vertices to all other vertices if directionality is ignored. A digraph is considered strongly connected if there are mutual paths between any two vertices in the digraph [163, 118].

For any graph, there need not be a path from one node to any other node in the network. Multiple sets of connected nodes, disjoint from each other, may emerge in a network. These sets of connected nodes are known as *components* of the network. Inherently, there are no paths from one set of connected components to another set of connected components. Within a directed graph, these may be considered weakly or strongly connected components. Further, in a digraph for a node v_i the set of all nodes from which there is a directed path to v_i, including the node itself is the *in-component* of v_i. Analogously, the set of all nodes that can be reach from a directed path from edges beginning at the nodes v_i, including v_i, is the *out-component* of v_i [163].

Network science often considers the largest connected component within a network. Many empirical networks inferred from data are known to have their largest connected component be a significant proportion of the network, also called the *giant component* [163]. For directed networks, these empirical networks can often be described by the "bow-tie structure", with a large strongly connected component, visualised as the central knot of the bow-tie, an in-component visualised as the left wing of the bow-tie, and an out-component visualised as the right wing of the bow-tie [163].

1.4.7 Latin Hypercube Sampling

Latin Hypercube sampling (LHS) is a statistical sampling method for generating random parameter values from a multi-dimensional parameter space. The LHS method divides the range of each variable in the parameter space into equal, non-overlapping intervals, with the number of intervals equal to the number of desired sampling points. A parameter value is then sampled at each interval for all variables ensuring that all intervals of each variable are represented in this sampling [126, 212].

As a practical two-dimensional example of the application of LHS, we suppose we want to study the effects of a drug on a population of female patients based on age and weight. We want to randomly sample the population using LHS. The range of

age of the patients is 15 – 65 years, and weights is 45 to 90 kilograms. We first select a number of desired sampling points, $n_{sampling} = 5$. We then divide the range of each variable into five equal intervals as follows. Note that, by convention of the numerical method, the upper bound of the last interval is included to cover the full range of the population.

Age: [15 – 25), [25 – 35) , [35 – 45), [45 – 55), [55 – 65]

Weight: [45 – 54), [54 – 63) , [63 – 72), [72 – 81), [81 – 90]

We then randomly select one value from each interval for both parameters.

Age samples: 21, 27, 40, 53, 62

Weight samples: 48, 58, 71, 76, 82

We then randomly permute without replacement the age and weight samples to form sampling points. This process ensures that the intervals of the variables are represented exactly once across the set of sampling points. The sampling points in this example are shown in Table 1.1 with the resulting LHS sampling visualised in Table 1.2, demonstrating that each row and column, pertaining to age and weight, contain exactly one sampling point. This is illustrative of the coverage of each variable's range provided by the use of LHS.

Sample Point	Age	Weight
Point 1	21	58
Point 2	27	76
Point 3	40	71
Point 4	53	82
Point 5	62	48

Table 1.1: LHS Sampling Points

Age / Weight	**45 – 54**	**54 – 63**	**63 – 72**	**72 – 81**	**81 – 90**
15 – 25		**X**			
25 – 35				**X**	
35 – 45			**X**		
45 – 55					**X**
55 – 65	**X**				

Table 1.2: LHS example visualisation. The interval ranges of the age and weight variables are shown in the rows and columns, with the **X** corresponding to sampling points used.

1.4.8 Biological Experimental Approaches

Scientists explore mechanisms and effects of biological functions and characteristics through a variety of approaches. Among these types of study methodologies are *in vitro* and *in vivo* studies [223, 239, 95]. *In vitro*, Latin for "in glass", experiments are those done out of a living organism, independent of the original biological environment. For example, experiments done in test tubes or Petri dishes are *in vitro* experiments [84, 109]. Studies performed *in vitro* allow for isolation and control of variables and experimental conditions. Additionally, *in vitro* studies skirt the need for the ethical considerations of testing on a live human or animal, especially in exploratory scientific phases. However, *in vitro* studies are limited in their ability to capture the complexity and interactions of the whole organism [223, 239, 95]. Another type of scientific investigation methodology is known as an *in vivo* study. *In vivo*, Latin for "in the living", experiments are those done inside of a living organism. This can refer both to animal or human organisms [84]. *In vivo* studies can more accurately capture the effects of biological functions and characteristics in the system and corresponding interventions. In particular, the many interactions occurring within the systems of living organisms are better represented. However, it is much more difficult to control variables in these types of studies. Furthermore, scientists must take into account the ethical considerations when conducting these experiments [223, 239, 95]. A comparison of *in vitro* and *in vivo* approaches is given in Table 1.3.

In the study presented in subsequent Chapters of this thesis, we consider experiments performed on human gut biopsy cultures. That is, samples are biopsied from the gut of human participants and then cultured outside of the body. This biopsy, specifically is considered *ex vivo*, meaning "outside of the body", which refers to biological material taken from a living organism and maintained outside of its body in an environment as close to the one in the organism as possible [84, 109]. However, the experimental perturbation study done on these culture models does not mimic the natural physiological environment and is considered an *in vitro* experiment. Hence, we refer to the experiment under consideration in subsequent Chapters as an *in vitro* experiment.

In general, as well as in cytokine analysis, experiments are undertaken in both *in vitro* and *in vivo* settings. *In vitro* experiments are particularly valuable when the goal is identifying or characterising very specific functions or behaviours of a biological entity. The controlled environments of the *in vitro* setting allow researchers to identify specific features of cytokines or response to cytokines. For example, using *in vitro* experiments, researchers can pinpoint the specific profile of cytokines that are elevated

after exposure to COVID-19 as well as distinguish the cytokine profile of those who suffered from long-COVID-19 [80]. The use of *in vitro* methods allows researchers to focus specifically on cytokine profiles in the response, while avoiding the ethical concerns of exposing individuals to the COVID-19 virus.

In vivo settings give a broader and more system-integrated perspective of the function of a cytokine. They allow researchers to capture how other immune regulators interact with the cytokine system. For example, *in vivo* studies in mice can show the effects of kidney removal and damage on cytokine production, clearance, and circulation, considering system-wide interactions [10]. However, these experiments are often more costly and require greater ethical consideration to implement. Both approaches are essential for developing a comprehensive understanding of a biological process, such as cytokine interaction, especially before progressing to clinical trials and therapeutic implementation [223, 239, 95, 84, 109].

Approach	Advantage	Disadvantage
In vitro	Isolation of variables Controlled environment Avoid ethical concerns More potential to standardise Less expensive	Missing biological features Lacks system-wide interactions
In vivo	Higher biological relevance System-wide incorporation	Ethical concerns Hard to isolate variables More expensive & time consuming

Table 1.3: Comparison of *in vitro* and *in vivo* experimental approaches.

Part I

The Extraction and Analysis of Cytokine Networks and Associated Dynamics in Inflammatory Bowel Disease

Chapter 2

Overview and Preliminaries

2.1 Overview

Biological systems are generally highly interconnected and complex. Their structure and interactions are crucial to understanding the emergence of biological phenomena with many biological systems having a highly interconnected nature [145, 182, 155, 106]. In particular, their fine-scale interactions give rise to emergent characteristics and mathematical modelling provides a framework for rigorously analysing such systems. [16]. Specifically, the regulation of the immune system is dynamically controlled in part by an intricate set of protein-receptor interactions, whose dysregulation can lead to autoimmune diseases, such as Inflammatory Bowel Disease (IBD). Central to the immune system are cytokines. These are secreted small signalling proteins that drive and regulate pro- and anti-inflammatory responses, controlling innate and adaptive immune responses, with a careful balance required to maintain capacity to drive immune responses without generating overt autoimmunity [128, 55, 38]. Given the essential role of cytokines in all aspects of immune responses, there is extensive interest in understanding how cytokines signalling pathways and networks regulate immune responses [224, 55, 32, 221, 72]. Cytokine therapeutics have been successfully applied as interventions in dysregulated signalling pathways to autoimmune diseases such as rheumatoid arthritis, Crohn's Disease, and multiple sclerosis, with many cytokine therapy medications on the market today [224].

Vital to the development of targeted cytokine therapies is an understanding of the interactions of cytokines. Given that a single cytokine may up- or down- regulate multiple other cytokines, including itself, the stimulation or inhibition of a single cytokine may have downstream effects through interactions with other cytokines that are not readily apparent [38]. This gives rise to a complex interaction structure of

multiple cytokines regulating other cytokines and themselves [72]. Given the complex interaction structure of cytokines, the cytokine system has been dubbed 'the cytokine network', a nomenclature describing the interactions of cytokines among themselves [72]. Typically, investigations of the structure of the cytokine network are undertaken preliminary through *in vitro* experiments measuring the expression of multiple cytokines given perturbations that are stimulatory or inhibitory in nature. Once preliminary insight into cytokine pathways is established by *in vitro* experiments, experimentation in mouse models is undertaken to describe more specifically if a cytokine is pro- or anti- inflammatory *in vivo*. These experiments may then be extended to the study of human biopsies [12] and clinical trials for the treatment of auto-immune diseases [224]. Through these methods, the role of a cytokine and its interactions with other cytokines is elucidated.

The structure of cytokine networks investigated via *in vitro* culture models is often described in a list-like structure, describing interactions between cytokines and their effect [12, 9, 72, 32]. These experiments give insight into the net effects of cytokines; however, the details of all possible interactions between any pair of cytokines are difficult to capture through overall effects and rarely reflect the complexity of biology observed *in vivo*. The task of classifying a cytokine network from data scales rapidly as the number of cytokines increases due to the combinatorial explosion in the number of possible cytokine interactions. Further, the determination of the response of a complex system is hindered without a full description of its intermediaries and an understanding of how these intermediaries affect the system behaviour [237, 70, 245], leading to the need for mathematical models that consider these multiple and interconnected cytokine interactions and resulting pathways in more detail.

The formalisation of the cytokine network as a mathematical network offers benefits to the understanding of cytokine interactions. Using a network, we can catalogue and visualise interactions between cytokines, as well as describe the strength of their interactions [36, 146]. One of the major advantages of a formal network is that one can impose dynamics on the network, describing how the network evolves in time [163, 120]. In this case, network nodes are cytokines and the connecting edges describe an interaction between these cytokines, such as association, up-regulation, or down-regulation. We may then impose ordinary differential equations on the network nodes, describing changes in cytokine concentration over time, thus giving a description of the propagation of a signal (or stimulus) across the network. Mathematical modelling using the networks approach provides a framework for capturing the complexity of the highly connected and non-linear interactions of cytokines. The

modelling of cytokine dynamics through network-based models also allows for predictions beyond the experimental setting, developing the productive cooperation between experimentation and modelling in the establishment of biological insights.

Within a network, structure and topology have a decisive influence on the resulting dynamics of the system [29, 209]. As such, there is a need for accurate network extraction methods applied to biological systems to correctly describe system connections and how within-system links influence behaviour, giving rise to biological phenomena [246, 182]. Several network extraction techniques have been successfully applied in biology, leading to broader understandings of biological systems as with the case of gene regulatory networks. However, many of these methods require large datasets in order to extract accurate networks [210, 142, 205, 70] or are unable to extract accurate networks when dealing with non-linear correlations between data in a dataset [245, 123], leading to an under-addressed deficiency in network extraction for systems with sparse data. Recent work from Jansen et al. [98] pertaining to network extraction from sparse datasets recovers a weighted and directed cytokine network describing cytokine interactions.

The remainder of this part of the thesis will be structured as follows. In Section 2.2 we outline the main concepts from which this work is derived. Specifically, we will describe the network extraction technique developed by Jansen et al. [98] and the model assumptions used to interpret the cytokine interaction network of interest in the previous literature. Motivated by previous literature elucidating cytokine systems from sparse human datasets, in Chapter 3 we theoretically investigate the properties and robustness of the dynamical system derived in the extraction of a cytokine interaction network from human gut biopsy cultures, focusing on the secretion of the cytokine IL-23, a key pathological cytokine associated with IBD. Noting cytokine clearance mechanisms feature *in vivo* but not in *in vitro* biopsy culture perturbation experiments, we additionally focus on using simulation and modelling to consider the impact of cytokine clearance as an initial investigation in theoretically replicating key *in vivo* features of the system.

2.2 State of the Art and Model Assumptions

In the following sections, we introduce the main concepts of the network extraction analysis technique from previous work by [98] from which this work builds upon. We will also explore the biological model assumptions used to analyse the cytokine interaction network under consideration in [98] and extracted from human gut biopsy

cultures from experiments by Aschenbrenner et al. [12]. We will analyse the dynamics of this extracted system in the study presented in Chapter 3.

2.2.1 Model Assumptions and Experimental Details

The system inherited from Jansen et al. [98] and Aschenbrenner et al. [12] is a model of cytokine network inferred from perturbation experiments of human gut biopsy cultures. The focus on the studies is on six key cytokines: TNFα, IL-6, IL-1α, IL-1β, IL-10, and IL-23 with particular attention paid to IL-23 due to indications of promising therapeutic effects upon its regulation. Stimulation during the experimental perturbations was given in all experimental conditions by lipopolysaccharise (LPS - a component of the bacterial cell membrane and a potent immune stimulator) and in one experimental condition by the addition of the cytokine IFNγ.

Experimental perturbations were undertaken by the blockage of cytokines via the introduction of an anti-cytokine blocker. This results in eight experimental conditions where stimulation by LPS is introduced in all perturbation conditions. These conditions are: a control condition of the stimulation of all cytokines with no anti-cytokine, a condition with the introduction of IFNγ with no anti-cytokine, and conditions of anti-TNFα, anti-IL-6, anti-IL-1α, anti-IL-1β, anti-IL-10, and a dual blockage condition of anti-IL-1α, anti-IL-1β. The model assumes the full blockage of the cytokine under its corresponding anti-cytokine experimental condition as well as no decay of cytokine in the *in vitro* setting. Further, the model also assumes the concentration of IFNγ is given only by its introduction into the experiment as a stimulator of the system, as the immune monocyte cells under consideration in the biopsy experiments are not known to produce IFNγ [98].

The model assumes the concentration of cytokines is represented by an ordinary differential equation of the form,

$$\frac{d[x]_i}{dt} = s_i \prod_u (1 + \alpha_{u,i}[x]_u(t)) \prod_v \frac{1}{(1 + \beta_{v,i}[x]_v)}, \tag{2.1}$$

where $[\cdot]$ denotes a cytokine concentration, i, u, v index over cytokines, $\alpha_{u,i}$ and $\beta_{v,i}$ are positive constants that indicate network edge weights denoting parameter values between a pair of cytokines, and s_i is the stimulating effect s_i provided by LPS. The first product models the activation of a cytokine via the influence of other cytokines, while the second product models the repression of the cytokine via the influence of other cytokines. The formulation of this Hill-like ODE system to describe the concentration of cytokines over time is given by [98].

2.2.1.1 Units of concentration

The measurement of the concentration of a cytokine i given in the dataset provided in [98] is expressed as the percentage of monocytes secreting i multiplied by the mean fluorescence intensity of these monocytes [204]. Fluorescence intensity gives a correlational measure between expression of molecules of interest and light emission [152]. Mean fluorescence intensity is non-dimensional, as is the percentage of monocytes, thus the units of concentration used in [98] and in this work are non-dimensional, reflecting relative changes from a fixed concentration scale.

2.2.2 Network extraction technique

The network extraction technique developed by [98] is intended for use in sparse datasets where other techniques are unlikely to be applicable. The methodology utilises maximum likelihood estimation and the Akaike Information Criteria to identify a set of possible network configurations and select a model that balances fit and complexity among possible configurations.

The obstacle to such a network extraction is scaling due to increases of computational complexity as the number of cytokines increases. Given a cytokine, the possible link connections considered to each cytokine, including itself, are an up-regulatory edge, a down-regulatory edge, and no edge. In the naive approach, for each cytokine n introduced, the combinatorial exploration of these links scales as 3^{n^2}. This quickly creates an intractable problem when each possible network produced is fit to data for larger datasets. To mitigate the increase of computational complexity, the methodology develops a procedure of reduction of possible network configurations to consider during network inference.

The procedure developed in [98] calculates the AIC of models in a subset of different possible configurations, given by

$$\mathrm{AIC} = D + 2p, \tag{2.2}$$

where p is the number of parameters or edges in the model and D is the distance to the data given by $D = \min_\theta \left(-2 \log \mathcal{L}(\theta)\right)$ (see Section 1.4.3). We note that the minimisation of the negative log-likelihood function, which gives us D, is computationally expensive. Using this distance D, the methodology identifies essential edges (to be defined) within a model of size p, considering all networks yet reducing the number of networks in which D must be explicitly calculated.

The method begins by considering only statistically significant edges in possible sub-configurations of the total number of possible edges. Supposing that there are N statistically significant edges, the method then generates a list of initial configurations of the network for every network size $1 \dots N$. The edges are introduced into the configuration according to their ranked statistical significance. For example, the initial configuration of size $p = 1$ contains the highest ranked statistically significant edge, the initial configuration of $p = 2$ contains the highest two ranked edges, and so forth. The distance D for each initial model configuration is calculated.

In order to consider whether there exists a configuration that is a better fit than the statistically significant edge initialisation for any size, the following procedure is implemented, which updates the selected model at each size until the best fit for each size is reached. Suppose that the set of all statistically significant edges is $\mathcal{O}$ and the initial configuration is S^p_{init}.

The procedure begins by defining the distance D of the initial configuration for a size p as the 'threshold' distance for that model size $D(S^p_{init}) = D^p_{thres}$. The procedure then considers the set of all statistically significant edges, $\mathcal{O}$, and a set of initially selected edges in the current model, S'.

1. Remove any combination of initially selected edges from the set of all statistically significant edges, $\mathcal{O}/S'$.

2. Calculate the distance of this configuration, $D(\mathcal{O}/S')$

3. If $D(\mathcal{O}/S') \geq D^p_{thres}$, then S' is considered 'essential'.

After edges are deemed 'essential', combinations of configurations with this set of edges removed need not be considered (as shown in [98]). By determining essential sets of edges, this methodology reduces the number of configurations where distances D must be explicitly calculated, reducing the computational cost of the search for the optimal model configuration to tractable levels.

After identifying the edge configuration that most reduces D for each size p, the AIC for each of these models is calculated and the model with the current lowest AIC is selected as the optimum model. With essential edges known, it is computationally feasible to do an exhaustive search during the network extraction process. We note that the minimisation of the log-likelihood is based on a deterministic trust region multi-start method approach in order to improve stability and avoid local minima. Initial guesses for the multi-start approach are based on Latin hypercube sampling. The implementation of the network extraction method developed by [98] is carried

out through the software Data2Dynamics, a MATLAB-based modelling environment utilised for parameter inference in dynamical systems, and recovers the topological structure of the network edges, as well as their associated parameter weights, $\alpha_{u,i}$ and $\beta_{v,i}$ [186, 185, 114].

Chapter 3

The Cytokine Network for IL-23 Regulation

3.1 Introduction

Cytokines are small diffusible signalling proteins that regulate immune responses, whose dysregulation is a major contributor to immune-mediated inflammatory diseases [221, 140, 72]. Treatment of autoimmune diseases, such as Inflammatory Bowel Disease (IBD), include the use of cytokine receptor blocking therapies including Ustekinumab (anti-IL12/23) and Infliximab (anti-TNF) to suppress dysregulated immune responses [160, 72, 224]. While there is an intense interest in this area, exemplified by Ustekinumab (Stelara) with sales of $10.4B for treatment of IBD, nonetheless the cytokine dynamics in human gut tissues are not well understood. Hence, there is extensive interest in investigating the structure and interactions of cytokine networks in tissue [145, 182].

In this chapter, we will describe a cytokine system that has been motivated by previous studies which captured the dynamics of cytokine interaction pathways, differentiating between direct and indirect interactions via intermediary cytokines, through the construction of a mathematical model of a cytokine network. We define direct cytokine effects as the interaction between two cytokines occurring independently of other cytokines. In contrast, indirect effects refer to the interaction between cytokines via an intermediary pathway. For example, suppose that cytokine A directly up-regulates cytokines B and IL-23, and cytokine B directly down-regulates IL-23. Then we say that cytokine A directly up-regulates IL-23 and indirectly down-regulates IL-23 (see Figure 3.1).

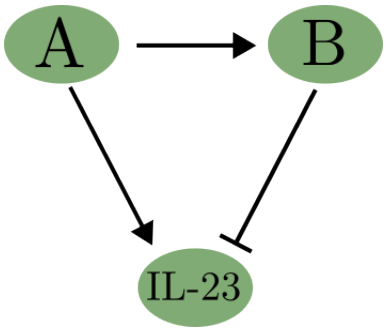

Figure 3.1: **Example of a simple network:** A simple network illustrating representations of cytokine interactions. Cytokine A up-regulates cytokines B and IL-23 directly, cytokine B down-regulates IL-23 directly. Cytokine A down-regulates IL-23 indirectly. The net effect of cytokine A is unclear and is also dependent on the presence of cytokine B.

As our focus and exemplar, recent work by [98] used an inference method to extract cytokine networks from *in vitro* data, extracting network topologies from sparse data derived from cytokine perturbation studies. The methodology was developed through application to data available from experimental perturbation studies of human gut biopsy cultures in the context of the cytokine IL-23 [12]. In particular, this study used a novel inference approach to fit a mathematical model consisting of a system of ODEs that describe the time evolution of each cytokine within a cytokine interaction network from biopsy culture data. The use of the ODE model captures how cytokine secretion evolves in time, enabling a quantification of the strength of interactions and regulatory nature of interaction between cytokines, and in principle allowing a description of the system dynamics. This methodology recovers the topology of the interacting cytokine system as a network, with cytokines as nodes and the interactions between them as weighted, directed edges, generating a prediction for the cytokine network, as given in Figure 3.2.

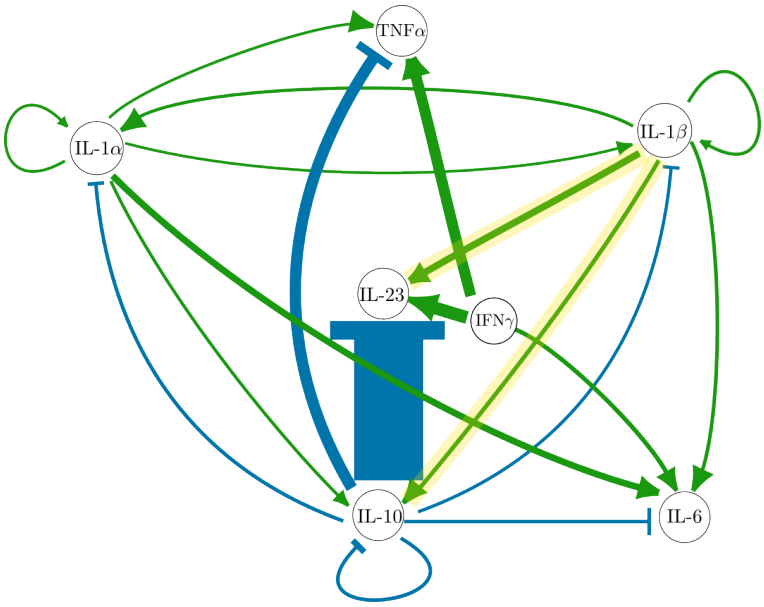

Figure 3.2: **Extracted Network from biopsy perturbation studies:** Pointed arrows denote up-regulatory effects from one cytokine to another, flat arrows denote down-regulatory effects, and the size of edges is representative of their respective parameter value for the strength of interaction between cytokines. In yellow, we highlight two predicted interaction pathways from IL-1β, an up-regulatory edge from IL-1β to IL-23 and an up-regulatory edge from IL-1β to IL-10, which in turn will down-regulate IL-23. These two predicted interactions of IL-1β create uncertainty as to its net effect on IL-23. We note that IFNγ acts only as a stimulatory agent in the given study and is not a cytokine of direct interest in the current context [98].

Hence, the methodology gives a description of the relationships and strength of interaction between key cytokines, as well as presenting a network inference technique that can be used to extract this information. However, there are aspects of this methodology, and its inferred network, that merit further investigation. Thus, our objectives are to firstly examine the impact of noise on the network inference process, and whether uncertainties in the data will fundamentally alter the network. This is relevant as experimental data is often noisy and it is unclear to what extent such experimental noise will impact the network inference, and also represents an important validation supplementing the use of methodology developed by [98] to perturbation studies with sparse data.

Secondly, we will examine the dynamics of the network to further understand the role of each of the driving cytokines in the described system on the regulation of the cytokine of interest, IL-23, as inherited from previous studies [12, 98], which only considered network extraction. We note that the methodology and available data emphasise an interest in the regulation of the cytokine IL-23, and we therefore

focus on predictions for IL-23 regulation. In the analysis of the dynamics of the system, we draw particular interest to the cytokine IL-1β, noting its influence on IL-23 is predicted to occur via multiple pathways in the extracted cytokine network, as highlighted in Figure 3.2 and its caption. Hence, we examine whether the role of IL-1β is to fundamentally up- or down- regulate IL-23. Further, given the complexity of cytokine networks, we also explore possible model reductions of our system.

Our final objective will be to consider the prospective translation of inferences from networks based on *in vitro* data to the *in vivo* setting. A fundamental difference *in vivo* is the rapid clearance of cytokines, which is not present in the biopsy culture studies. Hence, we examine whether inferred networks from *in vitro* data result in predictions that are insensitive to the inclusion of *in vivo* clearance rates and thus the extent to which predictions from models inferred via *in vitro* data may readily translate into the *in vivo* setting.

3.2 Methods

3.2.1 Quantitative summary of model

The cytokine network is inherited from [98] and based on the *in vitro* biopsy data presented in [12], with 6 cytokines of interest, TNFα, IL-1α, IL-1β, IL-6, IL-10, and IL-23, and one cytokine which is treated as a stimulatory agent, IFNγ. For experimental details beyond those given in this work, refer to [98] and [12].

In the experiments, monocytes were stimulated with lipopolysaccharide (LPS), an immunogenic protein typically found in the cell wall of bacteria, to provoke cytokine production. Several experimental conditions were undertaken, involving the blockage of one or more cytokines at a time, with cytokine concentration measurements given at $t = 16$ hours across experimental conditions. Under one experimental condition, the system was also stimulated by the cytokine interferon gamma (IFNγ), which is considered a stimulatory agent, as the monocytes in the experiment are not known to produce IFNγ [98, 73, 40, 30].

The resulting experimental data contains information regarding a complex series of interactions between cytokines as the monocytes are affected by the blockage of cytokine receptors in different experimental conditions. The data from the experiment pertaining to the extracted network assumes no decay of cytokine in the experimental setting [12].

The system is represented by model equations in the form,

$$\frac{d[x]_i}{dt} = s_i \prod_u (1 + \alpha_{u,i}[x]_u(t)) \prod_v \frac{1}{(1 + \beta_{v,i}[x]_v)}, \tag{3.1}$$

where $[\cdot]$ denotes a concentration, i, u, v index over cytokines, and $\alpha_{u,i}$ and $\beta_{v,i}$ are positive constants denoting the associated edge weights of parameter values between a pair of cytokines. In this context, the stimulating effect s_i is provided by LPS. The stimulatory effect of interferon gamma is represented in the model equations as the inclusion of the term $(1+\alpha_{5,1})$ where relevant, where the labels 5 and 1 denote the effect from IFNγ to IL-23.

Data from the *in vitro* biopsy experiments are fit to the mathematical model described in equation (3.1) and processed through the network inference technique detailed in [98]. The methodology is based on a deterministic trust region multi-start method, and the calculation of the Akaike Information Criteria (AIC) for prospective combinations of sub-networks [3]. The network extraction method is implemented using the software Data2Dynamics, a MATLAB-based modelling environment, recovering the topological structure of the network edges and their associated parameter weights, $\alpha_{u,i}$ and $\beta_{v,i}$ [186, 185, 114].

The full system (see Figure 3.2) inherited from [98] consists of 18 edges, including self-edges, with the direct network interactions of IL-23 stemming from the cytokines IL-10, IL-1β, and IFNγ. We note that the cytokines TNFα and IL-6 have only in-edges, and the system is dependent on the up- and down- regulatory effects of 3 cytokines of interest, IL-1α, IL-1β, and IL-10, as well as stimulatory effects from IFNγ. Thus, the dynamics of the system can be described by the up- and down-regulatory effects of IL-1α, IL-1β, IL-10, and IFNγ, which we term the regulatory cytokines. With the inclusion of IL-23 as the target of our investigation among the regulatory cytokines, we recover the cytokine network of interest that will be the subject of our investigation, shown in Figure 3.3.

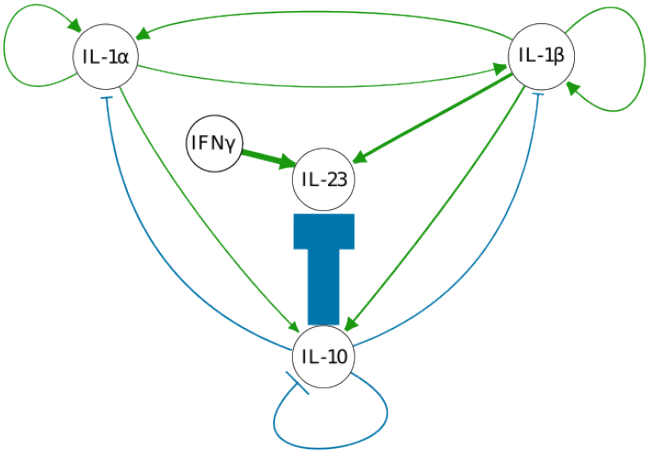

Figure 3.3: **Network of interest for IL-23:** The regulatory cytokines of the system and their relation to IL-23. Pointed green arrows show up-regulation and flat blue arrows show down-regulation. TNFα and IL-6 are excluded as the dynamics of the system is not dependent on their effect. Edge sizes are reflective of their relative parameter value with respect to each other.

3.2.2 Robustness to Noise

As the cytokine interaction network is inferred from experimental data, it is of interest to assess the robustness of the network extraction method to possible experimental noise. This is explored through the introduction of noise to the experimental biopsy data. Noise was added to the measurement of IL-23 across all experimental conditions as multiplicative noise in the form:

$$\text{noisy data} = (\text{original data})(1 + X), \tag{3.2}$$

where

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

Noise was introduced through 3 different noise distributions, low noise *i.e.* $\mathcal{N}(0, .05^2)$, medium noise *i.e.* $\mathcal{N}(0, .10^2)$, and high noise, *i.e.* $\mathcal{N}(0, .20^2)$. The prospect of inducing negative data is sufficiently rare so as not to occur in practice. For each given normal distribution, the noise was introduced to the original experimental data file to produce 100 generated noisy experimental data files per given distribution. The network extraction method was then implemented for each noisy experimental data

file to recover the corresponding now-noise-perturbed network.

We inspect the edges and edge weights of the recovered noise-perturbed network to determine whether the topology of the network was changed and how the introduction of noise influenced recovered parameter values. We then examine the relative difference in concentration of IL-23 at $t = 16$ hrs of each recovered perturbed network with respect to the original unperturbed network as a difference measure.

We note that multiple local minima, as can be found with the multi-start algorithm, means that the number of starts required to infer a network can vary. This can require constraints on initial start points, as further detailed by [98] and summarised in Appendix A.

3.2.3 The in vitro and in vivo systems

The *in vitro* system, as inherited from [98], is given by the following set of ordinary differential equations in the form of equation (3.1). For brevity of subscripts, each cytokine is given a numerical label, (see Table 3.1). As an example, $\alpha_{3,1}$ is read as an up-regulatory influence from IL-1β to IL-23, and $\beta_{2,1}$ is read as an down-regulatory influence from IL-10 to IL-23. The modelling equations are

$$\frac{\mathrm{d}[\text{IL-23}]}{\mathrm{d}t} = s_1(1+\alpha_{5,1})(1 + \alpha_{3,1}[\text{IL-1}\beta])\frac{1}{1 + \beta_{2,1}[\text{IL-10}])}, \tag{3.3}$$

$$\frac{\mathrm{d}[\text{IL-10}]}{\mathrm{d}t} = s_2(1 + \alpha_{4,2}[\text{IL-1}\alpha])(1 + \alpha_{3,2}[\text{IL-1}\beta])\frac{1}{1 + \beta_{2,2}[\text{IL-10}])}, \tag{3.4}$$

$$\frac{\mathrm{d}[\text{IL-1}\beta]}{\mathrm{d}t} = s_3(1 + \alpha_{3,3}[\text{IL-1}\beta])(1 + \alpha_{4,3}[\text{IL-1}\alpha])\frac{1}{1 + \beta_{2,3}[\text{IL-10}])}, \tag{3.5}$$

$$\frac{\mathrm{d}[\text{IL-1}\alpha]}{\mathrm{d}t} = s_4(1 + \alpha_{4,4}[\text{IL-1}\alpha])(1 + \alpha_{3,4}[\text{IL-1}\beta])\frac{1}{1 + \beta_{2,4}[\text{IL-10}])}, \tag{3.6}$$

Cytokine	Number label
IL-23	1
IL-10	2
IL-1β	3
IL-1α	4
IFNγ	5

Table 3.1: Number labels of cytokines

However, the introduction of cytokine clearance is necessary to approximate *in vivo* cytokine dynamics. *In vivo* cytokine clearance rates often vary widely, with many rates unknown due to difficulties of measurement, though is hypothesised that cytokines clearance half-lives are on the order of minutes [215].

Cytokine clearance is introduced to the ODE system via first order decay so that the modified ODE system becomes:

$$\frac{\mathrm{d}[\text{IL-23}]}{\mathrm{d}t} = s_1(1+\alpha_{5,1})(1+\alpha_{3,1}[\text{IL-1}\beta])\frac{1}{1+\beta_{2,1}[\text{IL-10}])} - \lambda_1[\text{IL-23}], \tag{3.7}$$

$$\frac{\mathrm{d}[\text{IL-10}]}{\mathrm{d}t} = s_2(1+\alpha_{4,2}[\text{IL-1}\alpha])(1+\alpha_{3,2}[\text{IL-1}\beta])\frac{1}{1+\beta_{2,2}[\text{IL-10}])} - \lambda_2[\text{IL-10}], \tag{3.8}$$

$$\frac{\mathrm{d}[\text{IL-1}\beta]}{\mathrm{d}t} = s_3(1+\alpha_{3,3}[\text{IL-1}\beta])(1+\alpha_{4,3}[\text{IL-1}\alpha])\frac{1}{1+\beta_{5,3}[\text{IL-10}])} - \lambda_3[\text{IL-1}\beta], \tag{3.9}$$

$$\frac{\mathrm{d}[\text{IL-1}\alpha]}{\mathrm{d}t} = s_4(1+\alpha_{4,4}[\text{IL-1}\alpha])(1+\alpha_{3,4}[\text{IL-1}\beta])\frac{1}{1+\beta_{2,4}[\text{IL-10}])} - \lambda_4[\text{IL-1}\alpha], \tag{3.10}$$

where λ_1, λ_2, λ_3, and λ_4 are the respective clearance rates. The introduction of clearance in the *in vivo* model gives rise to a steady state in the system. We explore the properties of this steady state through bifurcation analysis, inspection of the phase plane, and perturbations to the concentration of cytokines, taking interest in how the dependence of IL-23 on the other cytokines is impacted.

3.2.4 Sensitivity of IL-23

Given one of our aims is understanding the sensitivity of predicted dynamics, we conduct a global sensitivity analysis of IL-23 production to the regulatory cytokines of both the *in vitro* and *in vivo* systems. We conduct this global sensitivity analysis using eFAST, in MATLAB [143, 191].

3.2.5 Model Reduction

Network parameters, $\alpha_{4,3}$, $\alpha_{4,2}$ and $\alpha_{4,4}$, associated with the cytokine IL-1α are small (see Table 1, Appendix A) and sensitivity analysis will reveal a small impact on IL-23 production from IL-1α relative to other cytokines, hence we explore the reduction of the cytokine network model via the removal of IL-1α. We examine the difference in IL-23 concentration at $t = 16$ hrs between the original model and the reduced model

and investigate changes to the steady state when reducing the model in the *in vivo* system, by setting $\alpha_{4,3} = \alpha_{4,2} = \alpha_{4,4} = 0$, and no longer considering IL-1α.

In the reduced model, this reduced system is given as,

$$\frac{\mathrm{d}[\text{IL-23}]}{\mathrm{d}t} = s_1(1+\alpha_{5,1})(1+\alpha_{3,1}[\text{IL-1}\beta])\frac{1}{1+\beta_{2,1}[\text{IL-10}]} - \lambda_1[\text{IL-23}], \tag{3.11}$$

$$\frac{\mathrm{d}[\text{IL-10}]}{\mathrm{d}t} = s_2(1+\alpha_{3,2}[\text{IL-1}\beta])\frac{1}{1+\beta_{2,2}[\text{IL-10}]} - \lambda_2[\text{IL-10}], \tag{3.12}$$

$$\frac{\mathrm{d}[\text{IL-1}\beta]}{\mathrm{d}t} = s_3(1+\alpha_{3,3}[\text{IL-1}\beta])\frac{1}{1+\beta_{2,3}[\text{IL-10}]} - \lambda_3[\text{IL-1}\beta], \tag{3.13}$$

with the exclusion or inclusion of clearance differentiating between the *in vitro* and *in vivo* systems. We then examine the difference in IL-23 concentration at $t = 16$ hrs between the original model and the reduced model. We also investigate changes to the steady state when reducing the model in the *in vivo* system.

3.2.6 IL-1β, IL-10 and IFNγ

Finally, we consider the overall impact of IL-1β, IL-10, and IFNγ. We vary the concentrations of IL-1β and IL-10 between .001 and 10 units (for unit details see Appendix A) to investigate the effect on the concentration of IL-23 at $t = 16$ hrs *in vitro* and *in vivo*. We further explore the impact of the inclusion of these cytokines in the models of the *in vitro* and *in vivo* systems via the removal of IL-1β and IL-10 from the models. Finally, we look at the impact of IFNγ by removing the contribution of IFNγ to the stimulation value of IL-23 in the model equations, equivalent to setting $\alpha_{5,1}$ to zero.

3.3 Results

3.3.1 Robustness to Noise

Details of the impact of experimental noise sampled from three distributions, $\mathcal{N}(0, .05^2)$, $\mathcal{N}(0, .10^2)$, and $\mathcal{N}(0, .20^2)$, on the inference of the cytokine network are given below, with a summary in Table 3.2. In Figure 3.4, we show the percent relative difference in IL-23 concentration at the measurement time between the networks recovered from each noise-perturbed data file and the network recovered from the original data file. These relative differences are shown for low, medium, and high noise, with each cumulative distribution sorted into ascending order of relative difference to emphasise

the low differences in IL-23 concentration at the measurement time with respect to the original data in the vast majority of noise-perturbed files.

With the introduction of low noise to the experimental data, $\mathcal{N}(0, .05^2)$, 98 of the 100 generated noisy experimental data files produced the same qualitative network edges as the original data file. That is to say, an up- or down- regulatory edge between two cytokines recovered from the original non-perturbed file exists in the recovered network of the noise-perturbed file and vice versa. The remaining 2 of the 100 noisy files that do not produce the same network topology differ by containing an additional up-regulatory edge from IL-1α to IL-23. We measure the relative difference between the value of IL-23 at $t = 16$ hrs from the original non-perturbed network and each recovered noise-perturbed network. 98 of the 100 recovered networks, corresponding to those networks with the original network topology, have a relative difference in IL-23 concentration of less than 5%. The two recovered networks containing the additional edge from the cytokine IL-1α to the cytokine IL-23 give relative differences of IL-23 concentration of 17.6% and 21%.

With the introduction of medium noise, $\mathcal{N}(0, .10^2)$, 88 of the 100 generated noisy experimental data files produced the same qualitative network edges as the original data file. The remaining 12 of the 100 noisy experimental data files differ from the original network topology again by an additional edge from IL-1α to IL-23. Of the 100 recovered networks, 80 have a relative difference in IL-23 concentration with the original network at $t = 16$ hrs of less than 5%. 8 of the recovered networks have a relative difference in IL-23 concentration with the original network of between 5.2% and 6.6% at the measurement time. The 12 recovered networks with the additional edge from IL-1α to IL-23 give relative differences in IL-23 concentration of between 18.8% and 25.4%.

With the introduction of high noise, $\mathcal{N}(0, .20^2)$, 72 of the 100 generated noisy experimental data files produced the same qualitative network edges as the original data file. The remaining 28 of the 100 noise-perturbed experimental data files produced a network with an additional edge from IL-1α to IL-23. When measuring the relative difference between IL-23 concentration at $t = 16$ hrs produced by the original non-perturbed network and each recovered noise-perturbed network, 34 of the 100 recovered networks produced a relative difference in IL-23 concentration of less than 5%. Further, 30 of the 100 recovered networks produced a relative difference of between 5% and 10%, 10 of the recovered noise-perturbed networks produced a relative difference between 10% and 20%, and 26 of the recovered noise-perturbed networks produced a relative difference between 20% and 37.6%.

Sampled Noise Distribution	Networks Retaining Topology	IL-23 Difference <5%	IL-23 Difference of Remaining Networks
$\mathcal{N}(0, .05^2)$	98	98	17.6% and 21%
$\mathcal{N}(0, .10^2)$	88	80	8 networks between 5.2% and 6.6% 12 networks between 18.8% and 25.4%
$\mathcal{N}(0, .20^2)$	72	34	30 networks between 5% and 9.9% 10 networks between 10% and 20% 26 networks between 20% and 37.6%

Table 3.2: **Summary of Robustness to Noise:** Sampled noise distributions of experimental noise and their impact on the recovery of the cytokine network topology and concentration of the cytokine of interest IL-23. We show how many networks, of 100 recovered networks, retain the network topology recovered from unperturbed experimental data. We also consider the impact of introduction of noise on the concentration of IL-23 by showing the number of networks with a difference from the unperturbed IL-23 concentration of under 5% in the third column, as well as the range of percent difference with the unperturbed concentration of IL-23 of the networks recovered from noise that do not meet the under 5% criteria in the fourth column.

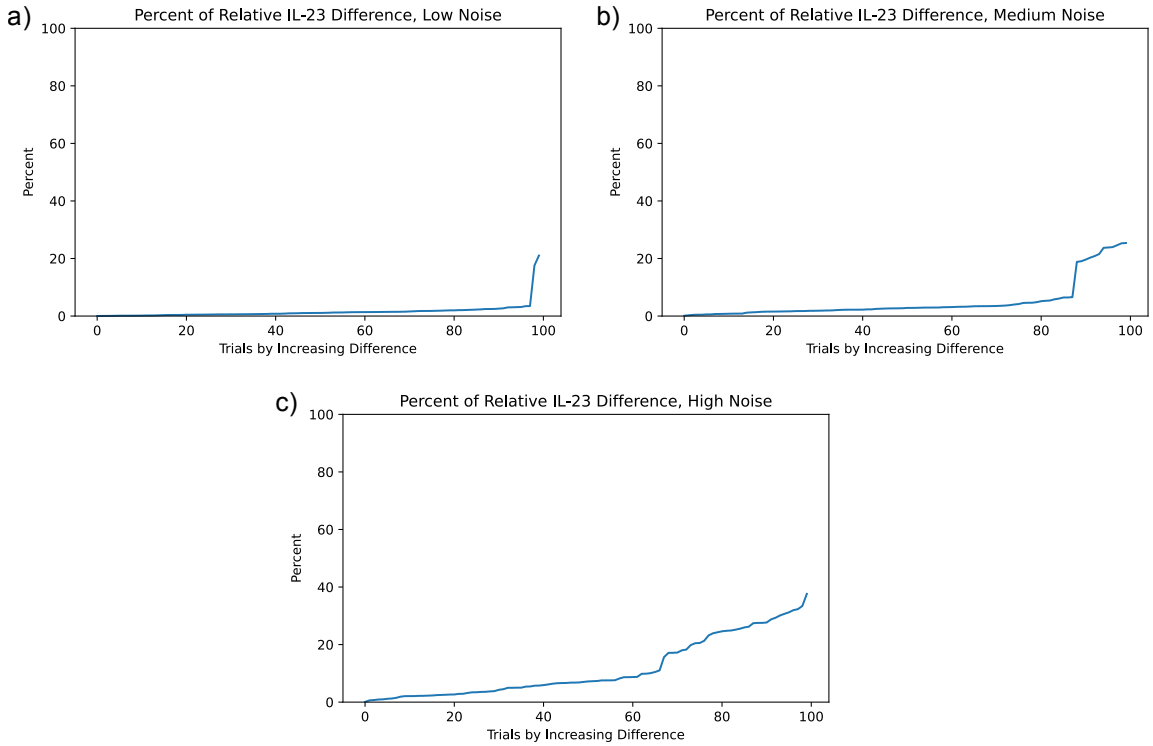

Figure 3.4: **Relative difference in IL-23 concentration between recovered networks based on noisy data files and the original experimental data:** The three different noise distributions introduced into the experimental data are represented. The percent relative difference of the concentration of IL-23 at measurement time between a noisy recovered network and the experimentally recovered network is shown for each of the 100 trials taken from a noise distribution, with low noise shown in panel **a)**, medium noise shown in panel **b)**, and high noise shown in panel **c)**. In spite of the introduction of relatively large noise to the experimental data, the relative difference in concentration of IL-23 at measurement time remains generally low.

3.3.2 The Steady State

The balance between production and cytokine clearance in the modelled *in vivo* system gives rise to a single steady state. For the numerical values explored computationally, the system exhibits this steady state as a globally attracting basin. Bifurcation analysis demonstrates that this steady state undergoes no bifurcations as the half-lives of cytokine clearance are varied within presumed *in vivo* values, while the steady state value for IL-23 decreases monotonically with clearance as shown in Figure 3.5, highlighting the difference between *in vitro* and *in vivo* model prediction. The cytokine clearance constant is equal to $\ln 2/t_{1/2}$ where $t_{1/2}$ is the half-life of cytokine measured in hours.

The predicted concentration of cytokines over time *in vitro* and *in vivo* are shown in Figure 3.6a)b), where we can clearly see the emergence of a steady state. We observe that through the introduction of clearance and considering half-lives to be on the order of minutes, the system is predicted to rapidly approach its fixed-point equilibrium (see Figure 3.6). Even with half-lives considered long *in vivo*, the system still reaches its equilibrium quickly with respect to the measurement time. This is elaborated in Figure 3.6a)b), which shows the concentrations of regulatory cytokines over time at different half-lives. Indeed, a characteristic timescale $\tau = 1/\lambda$ where cytokine clearance half life is 5 minutes, that is $\lambda \approx \ln 2/(0.0833)$ hrs, gives a ratio of the measurement timescale to relaxation timescale of

$$\frac{\tau_{measure}}{\tau_{char}} \approx \frac{16}{.12} \approx 133, \tag{3.14}$$

showing that the *in vivo* system is predicted to relax to its steady state two orders of magnitude faster than the measurement timescale. Even at 45 minute cytokine half-life clearance, which is considered long with respect to potential assumed half-life values, this ratio remains above an order of magnitude. We therefore introduce the quasi-steady-state assumption. Considering equations (3.7) − (3.10), and imposing the quasi steady state relation $\mathrm{d}[\text{IL-1}\alpha]/\mathrm{d}t \approx 0$, we recover expressions for IL-10 and IL-1β nullclines in terms of each other and recovered parameter values. We can represent the system in terms of IL-10 and IL-1β, and thus explore the steady state in terms of the phase plane geometry in terms of the concentrations of IL-10 and IL-1β in Figure 3.6c)d). Here we show the phase portraits at two different half-life values, 5 minutes and 45 minutes. We notice that the trajectory of the system closely follows the IL-10β nullcline when the half-life is 5 minutes, and reaches its steady-state quickly. Likewise, in spite of half-lives considered long *in vivo*, the system still reaches its equilibrium quickly with respect to the measurement time. This is further elaborated in Figure 3.6a)b) showing the concentrations of regulatory cytokines over time at different half-lives.

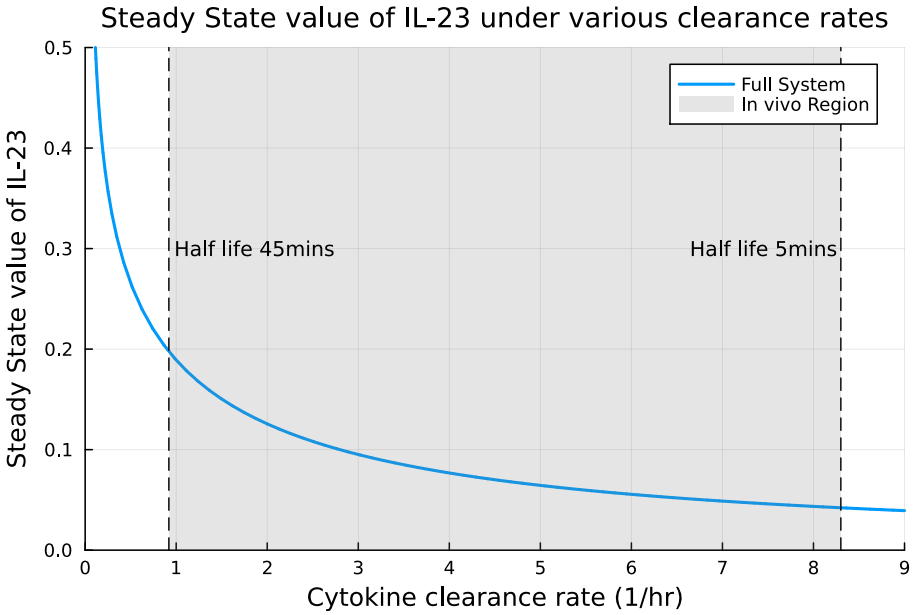

Figure 3.5: **The steady state value of IL-23 as clearance rates are varied:** As cytokine clearance rates are increased, the steady state value of IL-23 decreases monotonically. The clearance rate is given by $\lambda = (\ln 2)/t_{1/2}$ where $t_{1/2}$ half-life in hours. The shaded region indicates the *in vivo* region, with the values of the half lives at 5 and 45 minutes indicated.

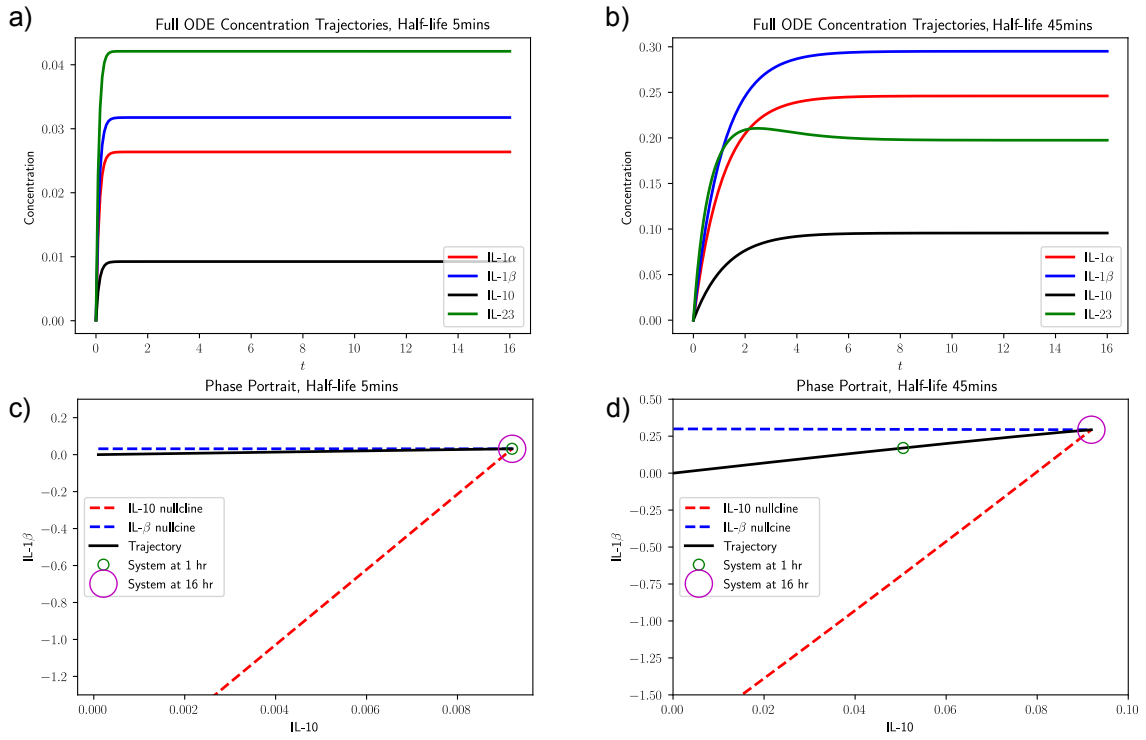

Figure 3.6: **Phase Portraits *in vivo*: a)** Predicted concentration of cytokines from modelling equations at 5 minute cytokine clearance. **b)** Predicted concentration of cytokines from modelling equations at 45 minute cytokine clearance. **c)** Phase portrait of IL-10 and IL-1β at a 5 minute half life. The state of the system is indicated at 1hr (green circle) and at the measurement time of 16hrs (purple circle). **d)** Phase portrait of IL-10 and IL-1β at a 45 minute half life. The state of the system is indicated at 1hr (green circle) and at the measurement time of 16hrs (purple circle). We note the system rapidly approaches its steady state at physiologically relevant half-lives, and this stability is maintained even when clearance is extended beyond predicted half-life values.

3.3.3 Sensitivity analysis

To determine the influence of the regulatory cytokines on IL-23, we conduct a global sensitivity analysis using eFAST on the *in vitro* and *in vivo* systems. The sensitivity analysis gives us a preliminary indication of the role of each cytokine and suggests directions from which to proceed as we consider our analysis of the system, focusing our subsequent attention on the effects of cytokines associated with high relative parameter sensitivity. The explicit values for S_{Ti} and S_i above our selected threshold, *i.e.* values greater than 1×10^{-2}, for the *in vitro* and *in vivo* systems are given in Tables 3.3 and 3.4, respectively. We chose 1×10^{-2} as the threshold for sensitivity values as parameters with sensitivity indices smaller than this contribute little to the

overall impact on IL-23 concentration, particularly when compared to unity.

For the global sensitivity analysis of the *in vitro* system via eFAST, we set the number of desired samples per search curve, NS, equal to 500 and the number of resampled search curves, NR, equal to 17 with a total of 16 parameters and one dummy parameter. The number of parameters result from the number of edges, the effect of LPS stimulation on each cytokine, and a dummy parameter. Sensitivity is measured between 0 and 1 in practice, though it is possible for S_{Ti} to exceed unity. Of the 16 parameters, 6 parameters have a total order sensitivity index above the threshold of 1×10^{-2} and 5 parameter have a first-order sensitivity index above threshold. The values of S_{Ti} and S_i for the dummy parameters are 3×10^{-3} and effectively zero, respectively.

In the *in vivo* system, we set NS equal to 250 and NR equal to 20, with a total of 21 parameters after the inclusion of clearance constants and the dummy parameter. Of the 20 parameters, 9 parameters have a total order sensitivity index above threshold and 8 parameter have a first-order sensitivity index above threshold. The sensitivities of S_{Ti} and S_i for the dummy parameters are 4×10^{-4} and zero, respectively.

To assess parameter impact, in the estimation of S_{Ti} and S_i, eFAST varies each parameter within a given range. In this investigation, the parameters s_i, $\alpha_{u,i}$, and $\beta_{v,i}$ vary between the lower and upper bounds given in Table 1. The explicit values for S_{Ti} and S_i above threshold, *i.e* values greater than 1×10^{-2}, obtained from the implementation of eFAST for the *in vitro* and *in vivo* systems are given in Tables 3.3 and 3.4.

In vitro, IL-23 production is highly sensitive to the stimulatory agents of the experiment, LPS and IFNγ, as expected. eFAST global sensitivity analysis also reveals that for the regulatory cytokines of interest, IL-23 production is most sensitive to changes in parameter value of direct links emanating from the cytokines IL-10 and IL-1β. The sensitivity to IL-10 is twice that of IL-1β in the *in vitro* system. The sensitivity of parameters associated with IL-1α are negligible (see Figures 3.7 and 3.9). These results suggests that we may be able to influence IL-23 through the manipulation of IL-10 and IL-1β *in vitro*.

Global sensitivity analysis of the *in vivo* system reveals that the sensitivity of IL-23 is dominated by cytokine clearance rates and the stimulatory effects of LPS and IFNγ (see Figures 3.7 and 3.8). Stimulatory effects are included in Figure 3.7 and normalised to unity in Figure 3.8. The normalisation to unity after removal of stimulatory effects more clearly shows the effect on IL-23 from the up- or down-regulatory links of other cytokines.

As in the *in vitro* system, IL-23 is sensitive to direct links from IL-10 and IL-1β *in vivo*, and insensitive to edges to or from IL-1α (see Figures 3.7 and 3.9). With the inclusion of cytokine clearance, the *in vivo* sensitivity to edges from IL-10 and IL-1β is reduced. Additionally the sensitivity to the edge from IL-10 to IL-1β that is present in the *in vitro* system is reduced in sensitivity beyond our threshold value for sensitivity *in vivo* (see Figure 3.9). Nonetheless, while the sensitivities and thus scale of responses fundamentally differ between *in vitro* and *in vivo* models, we note that the *in vivo* model has inherited some important features of the *in vitro* model. In particular, the *ranking* of both the first and total sensitivities are the same for both models on considering only the parameters that are in both models and above a small sensitivity threshold of 10^{-2} (see Tables 3.3 and 3.4).

Parameter	First Order Sensitivity	Total Order Sensitivity	Ranking by S_i	Ranking by S_{Ti}
s_1	0.3373	0.4378	1	1
s_2	0.0336	0.0522	5	5
α_{31}	0.0677	0.1058	4	4
α_{32}	$< 1 \times 10^{-2}$	0.0106	6	6
α_{51}	0.2010	0.2734	2	2
β_{21}	0.1921	0.2680	3	3
dummy	0.0000	0.003	-	-

Table 3.3: **Sensitivity *in vitro*:** First Order Sensitivity, S_i, and Total Order Sensitivity, S_{Ti} of IL-23 to *in vitro* model parameters. Larger values indicate higher sensitivity to a model parameter. For the *in vitro* model system, IL-23 production is most sensitive to parameters associated with LPS stimulation, IFNγ, and IL-10.

Parameter	First Order Sensitivity	Total Order Sensitivity	Ranking by S_i	Ranking by S_{Ti}
λ_1	0.2731	0.4139	1	1
λ_2	0.1306	0.2049	5	5
λ_3	0.0129	0.0290	4	4
λ_4	0.0114	0.0254	2	2
s_1	0.2207	0.3374	3	3
s_2	$< 1 \times 10^{-2}$	0.0142	-	-
α_{31}	0.0299	0.0562	-	-
α_{51}	0.0857	0.1690	-	-
β_{21}	0.0520	0.0905	-	-
dummy	0.0000	0.0004		

Table 3.4: **Sensitivity *in vivo*:** First Order Sensitivity, S_i, and Total Order Sensitivity, S_{Ti} of IL-23 to *in vivo* model parameters. Larger parameter values indicate higher sensitivity to a model parameter. For the *in vivo* model system, IL-23 production is most sensitive to clearance rates, parameters associated with LPS stimulation, IFNγ, and IL-10. Excluding clearance rates, as these are not present in the *in vitro* model, the ranking of sensitivity is the same as the *in vitro* model for parameters that exceed threshold for both the *in vivo* and *in vitro* models (see Table 3.3).

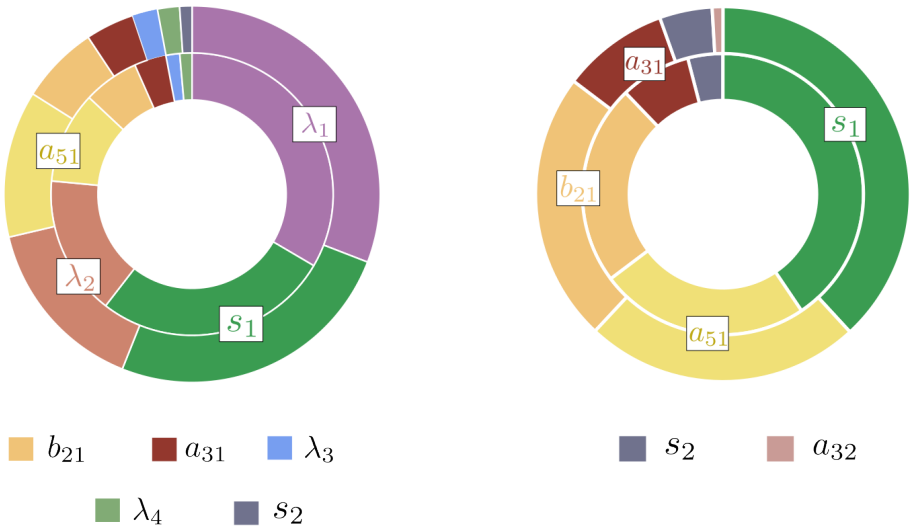

Figure 3.7: **Sensitivity analysis in proportion:** The sensitivity of IL-23 concentration at measurement time to system parameters for the *in vivo* system, left, and the *in vitro* system, right. The outer ring shows the total order sensitivity index, S_{Ti}, and the inner ring the first-order sensitivity index, S_i. The arc segment of each ring subsection is proportional to the sensitivity index of the parameter indicated. That is to say, the first order inner ring (measured 0 to 1) the arc segment is equal to the corresponding angle over 2π. The total order outer ring arc segment as the fraction of all arc segments is equal to the corresponding angle over 2π. Sensitivity indices only above 10^{-2} are depicted. λ_i denotes cytokine clearance and s_i denotes stimulatory effects on cytokine i. Cytokines IL-23, IL-10, IL-1β, IL-1α, and IFNγ are labelled 1-5, respectively (see Table 3.1).

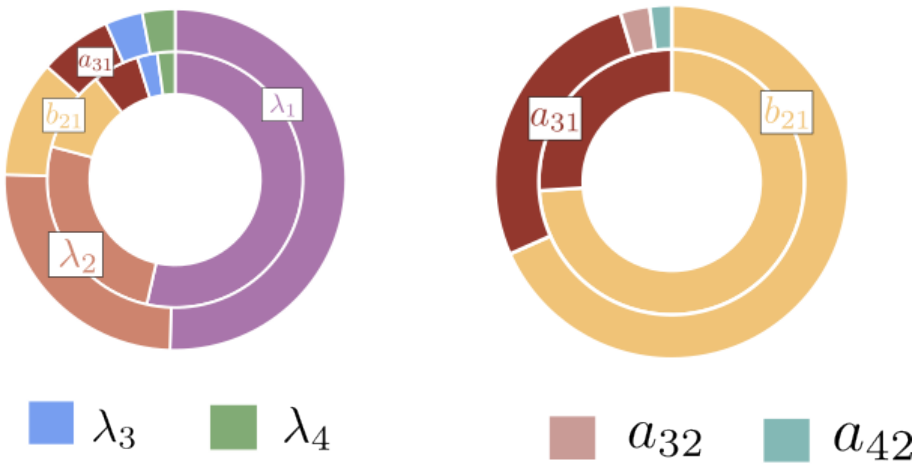

Figure 3.8: **Normalised sensitivity analysis:** The sensitivity of IL-23 concentration at measurement time to system parameters, normalised to unity once the stimulatory agents (LPS, IFNγ) have been removed. The *in vivo* system is shown on the left, and the *in vitro* system is shown on the right. The outer ring shows the total order sensitivity index, S_{Ti}, and the inner ring the first-order sensitivity index, S_i. Ring arc segment proportions are inherited as in Figure 3.7. Only edges with total order sensitivity above 10^{-2} are shown. λ_i denotes cytokine clearance and s_i denotes stimulatory effects on cytokine i. Cytokines IL-23, IL-10, IL-1β, IL-1α, and IFNγ are labelled 1-5, respectively (see Table 3.1).

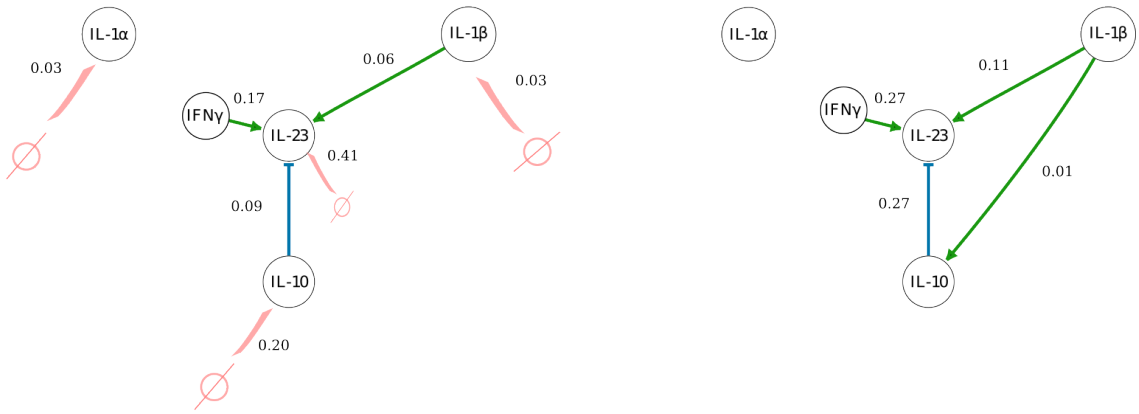

Figure 3.9: **Network description of total order sensitivity analysis:** A visual representation of the network of interest with edge labels showing the total order sensitivity of the global sensitivity analysis for the most important parameters. Left, *in vivo* system, with pink arrows representing cytokine clearance, again with edge weights giving the total sensitivity. Right, *in vitro* system. Only edges with total order sensitivity above 10^{-2} are shown.

We note that the relative simplicity of the networks derived from the eFAST sensitivity analysis of cytokine interactions means that the model predicts on a small number

of feasible changes, which can in turn have a significant impact on the concentration of IL-23.

3.3.4 Model Reduction via removal of IL-1α

We explore model reduction through the removal of IL-1α, as the sensitivity to parameter values associated with IL-1α is small relative to links from other regulatory cytokines.

We remove IL-1α from the *in vitro* and *in vivo* system models and examine the concentration of IL-23 at $t = 16$ hrs. In the *in vitro* system model, we find a relative difference in concentration of IL-23 at measurement time of 10.2% upon full removal of IL-1α from the system, despite its small edge weights in the original network. While our experimental data is constrained to a measurement time at $t = 16$ hrs, we infer from the model that the relative difference between the full model and the model with IL-1α removal increases as measurement time is increased in the *in vitro* model (see Figure 3.11).

In the *in vivo* model, the relative concentration change of IL-23 after the removal of IL-1α is .07% when the half-life of cytokine clearance is 5 minutes. This relative differences reaches 3.7% when clearance half-life reaches 1 hour. This is indicative of a more accurate model reduction of the *in vivo* system model in comparison to the *in vitro* system model. Further, in the *in vivo* model under the removal of IL-1α, the steady state value of IL-23 remains essentially the same for smaller half-lives of cytokine clearance, which explains the smaller discrepancies for IL-1α removal in the *in vivo* model (see Figure 3.10). This reinforces that the means of altering the system are small, in particular that manipulating IL-1α is predicted to have minimal effect. Concentration differences at the measurement time between this model reduced system and the full system are shown in Figure 3.11 in both the *in vivo* and *in vitro* model settings. This figure illustrates the minor discrepancies that arise in the predicted concentration trajectories of IL-23 under the *in vivo* model predictions across the range of considered half-life values. These discrepancies become markedly larger in the *in vitro* model prediction, emphasising the difference between the validity of the model reduction under *in vivo* and *in vitro* model assumptions.

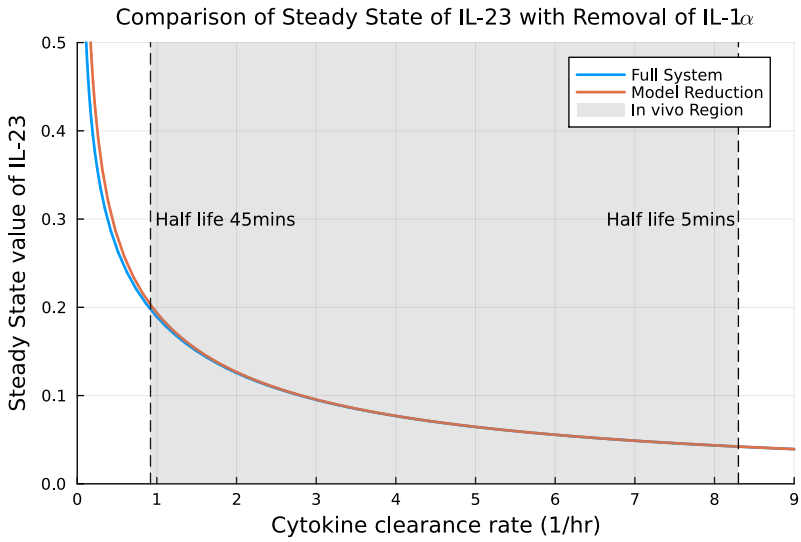

Figure 3.10: Comparison of IL-23 steady state value between full system and model reduction with removal of IL-1α as clearance rates are varied, highlighting that the impact of IL-1α is weak. Clearance rates are shown with half lives varied, with longer half lives being represented closer to the y-axis. The decay rate is given by λ where $\lambda = (\ln 2)/t_{1/2}$ and $t_{1/2}$ is half-life in hrs.

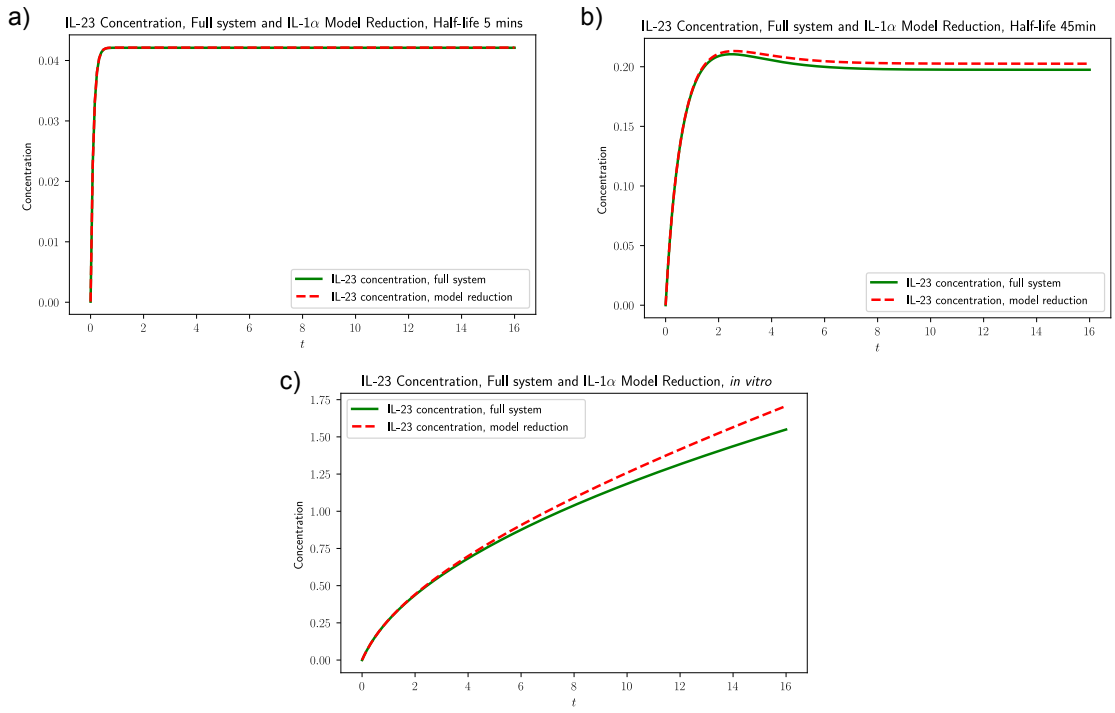

Figure 3.11: **Comparison of IL-23 concentration with IL-1α model reduction: a)** The predicted concentration of IL-23 according to the *in vivo* model shown in the full system (solid green) and the model reduced system without consideration of IL-1α (dashed red) with a half life of 5 minutes. **b)** The model predicted concentration of IL-23 according to the *in vivo* model in the full system and the model reduced system without consideration of IL-1α with a half life of 45 minutes. **c)** The model predicted concentration of IL-23 according to the *in vitro* model in the full system and the model reduced system without consideration of IL-1α. We see that as half lives are increased, the predicted difference in IL-23 concentration increases. In the *in vitro* system, the model reduction predicts more significant concentration difference in IL-23 at measurement time.

3.3.5 The effect of IL-1β

In this section, we investigate the overall effect of IL-1β on IL-23 in both the *in vitro* and *in vivo* models. As previously mentioned, whether IL-1β up- or down-regulates IL-23 production is not clear from the network representation (see Figure 3.3) nor in the model equations (3.3) − (3.6) and (3.7) − (3.10) because of the interplay between IL-1β and IL-10. Differences may also arise between the *in vitro* and *in vivo* system models due to the introduction of clearance in the latter.

We first inspect the influence of IL-1β concentration on IL-23 production. In the *in vitro* model, we find that the effect of IL-1β concentration is strictly up-regulatory. IL-23 concentration as a function of initial IL-1β concentration is monotonic, with a

decelerating gradient of response. In the *in vivo* model, the effect of IL-1β concentration has little to no effect on the concentration of IL-23 at $t = 16$ hrs (Figure 3.12), as possible impacts are mitigated by cytokine clearance.

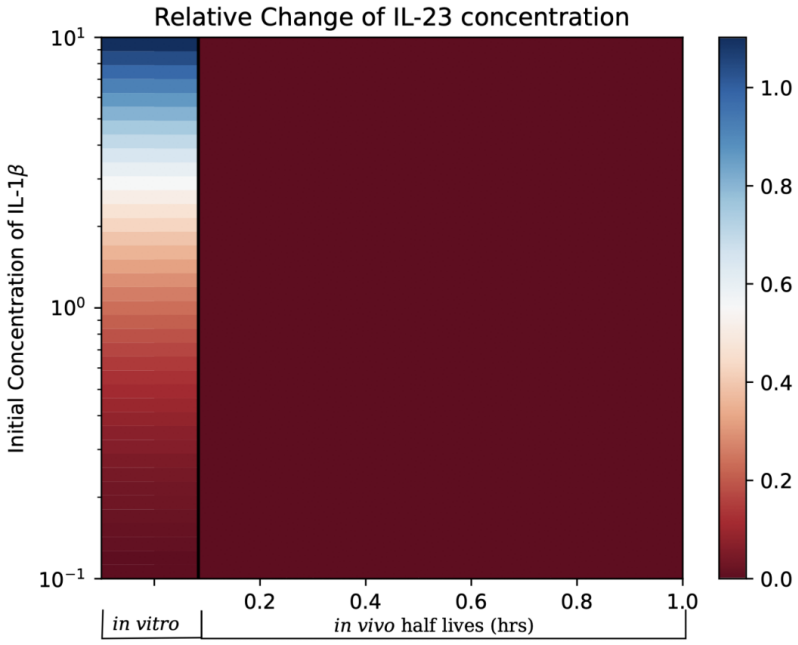

Figure 3.12: **Relative change in IL-23 concentration as initial concentration of IL-1β is varied. *In vivo* half-lives imply IL-23 is insensitive to changes in initial IL-1β across orders of magnitude, as is visually represented by the absence of a colour gradient. This contrasts with the effect of altering initial IL-1β *in vitro*, where clearance is not present in the model**. The *in vitro* system model with no clearance is shown on the left, and the *in vivo* system model on the right, with various half-life values for clearance. We see that for the *in vitro* model, increasing the initial concentration of IL-1β has an up-regulatory effect on the concentration of IL-23 at measurement time. However, regardless of the range of clearance rate in the *in vivo* model, the concentration of IL-23 at measurement time does not undergo noticeable change as initial concentration of IL-1β is varied.

We also consider the effect of the full removal of IL-1β from the system, to further investigate the influence of IL-1β on IL-23 production. In the *in vitro* model, we find large relative differences between the full model and the model where IL-1β has been removed, with a relative reduction in IL-23 production of approximately 50% at $t = 16$ hours. Hence, we see the heavy up-regulatory influence of IL-1β in the *in vitro* system (see Figure 3.14c)).

In the *in vivo* system model, we find differences in the steady state value of IL-23 with the extracted edge weights and clearance rates used previously under the removal of IL-1β from the system (see Figure 3.13 and Figure 3.14a)b)). We can see that the

inclusion of the IL-1β cytokine in the model equations has a clear up-regulatory effect on IL-23 production and this effect varies depending on the half-life of cytokine clearance. The trajectories of IL-23 concentration are shown in Figure 3.14 for *in vivo* and *in vitro* model settings, emphasising model prediction differences with even small half-life clearance values. This contrasts with results from model reductions of IL-1α where the model reduction has validity for small cytokine clearance values.

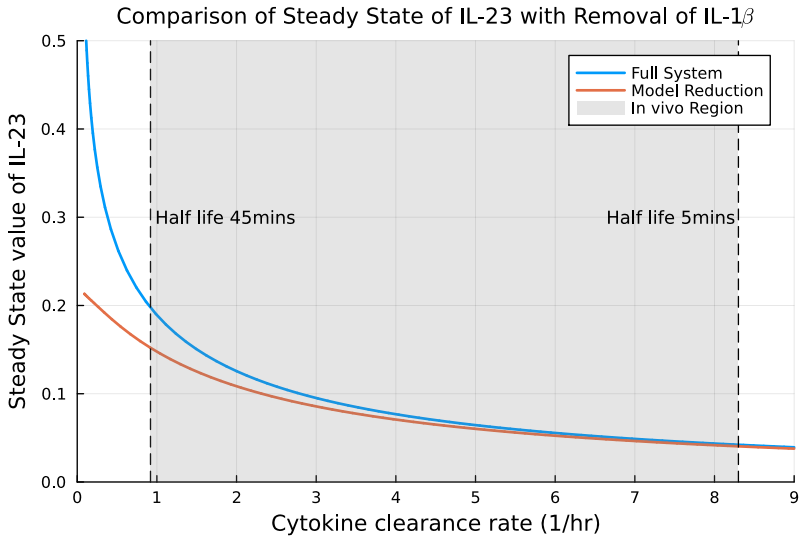

Figure 3.13: Comparison of IL-23 steady state value between full system and model reduction with removal of IL-1β as clearance rates are varied. Clearance rates are shown with half lives varied noting the clearance rates are given by λ where $\lambda = (\ln 2)/t_{1/2}$ and $t_{1/2}$ half-life in hrs.

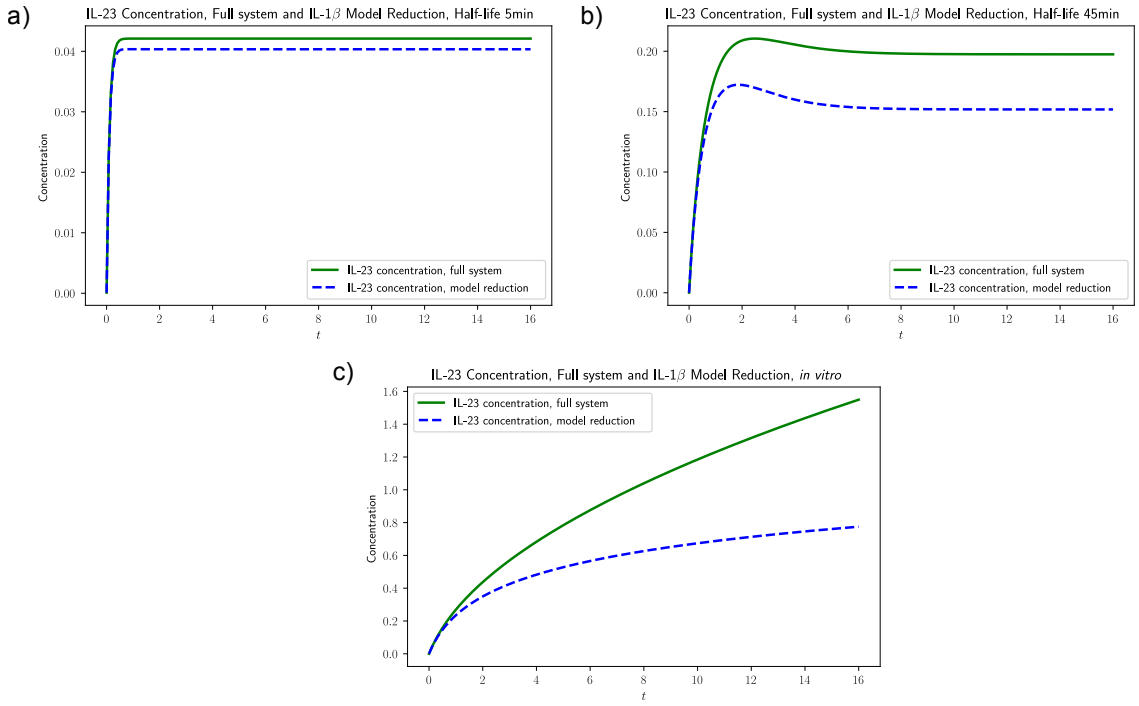

Figure 3.14: **Comparison of IL-23 concentration with IL-1β model reduction: a)** The model predicted concentration of IL-23 according to the *in vivo* model shown in the full system (solid green) and the model reduced system without consideration of IL-1β (dashed blue) with a half life of 5 minutes. **b)** The model predicted concentration of IL-23 according to the *in vivo* model in the full system and the model reduced system without consideration of IL-1β with a half life of 45 minutes. **c)** The model predicted concentration of IL-23 according to the *in vitro* model in the full system and the model reduced system without consideration of IL-1β. We see that similar to the model reduction of IL-1α, as half lives are increased, the predicted difference in IL-23 concentration increases. However, the model reduction of IL-1β shows greater differences in IL-23 concentration at measurement time across half-live values. In the *in vitro* system, the model reduction predicts large relative concentration difference in IL-23 at measurement time.

3.3.6 IL-10 and IFNγ

The effects of IL-10 are strictly down-regulatory, as expected and aligning with generally accepted previous literature [12, 113]. Increasing the initial concentration of IL-10 causes considerable down-regulatory effects both *in vitro* and *in vivo* for model predictions. When IL-10 is removed from the system, the concentration of IL-23 increases drastically.

IFNγ is a stimulatory agent under the experimental conditions. Removal of IFNγ

from the *in vitro* or *in vivo* system model decreases the value of IL-23 concentration at measuring time. All other qualitative results on the systems, such as the existence of a steady state and the possibility of an *in vivo* model reduction, remain the same.

3.4 Discussion

Manipulating and controlling cytokines is of intense interest, but their highly networked interactions entail that studies of cytokine regulation are difficult. In turn, this motivates the use of mathematical modelling in characterising the dynamics and behaviour of cytokine systems of interest. In particular, data from cytokine cultures of patient biopsies has been used to suggest how cytokines interact with the cytokine IL-23 in order to identify upstream regulators of IL-23 secretion control. The extraction and construction of a cytokine network pertaining to IL-23 has been explored based on gut biopsy data [98]; however the dynamics of the recovered cytokine network have not yet been investigated. Further, the robustness of the network extraction method to experimental noise and the inclusion of cytokine clearance as an *in vivo* extension of the model remained to be considered. Hence, our objectives have been to understand the dynamics of the *in vitro* system predicted from the biopsy culture data, as motivated by this IL-23 example, to assess the robustness of the network extraction method to noise, and to explore the impact of *in vivo* clearance rates on model predictions. In our analysis, we have first explored the robustness of the methodology by attempting to recover the cytokine interaction network given in [98] under three conditions of experimental noise. This assessment is important as experimental cytokine data is often noisy, and gauging robustness is instrumental for evaluating this network recovery methodology for the sparsity-characterised cytokine interaction data presented in [98]. We proceeded to explore the dynamical behaviour that arises from the cytokine interaction network according to the ODE system extracted as part of the network determination, along with the sensitivity of this dynamical behaviour to perturbations and model reductions. We have also explored how the features of the *in vitro* system model may translate into *in vivo* features, and if the predictions of the *in vitro* system model are reflected in the *in vivo* system model via extending to a simple model of the *in vivo* system by introducing the clearance of cytokines. The introduction of clearance gives rise to a simple dynamical system that globally relaxes to a single steady state. However, details regarding the features of this system are important in the determination of the sensitivity of IL-23 to other regulatory cytokines.

We began by showing that the network extraction method for the *in vitro* system, from which the *in vivo* system model will also be generated, is generally robust to experimental noise in the measurement of IL-23. We inspected the network recovery method under low, medium, and high random perturbations. As the amount of noise was increased, an additional up-regulatory edge between IL-1α and IL-23 can be predicted from the network extraction. The inclusion of this edge has an up-regulatory effect on the concentration of IL-23 at measurement time, though the relative magnitude of the effect is small. This additional edge was predicted more frequently as noise was increased and indicates that the direct effect of IL-1α may be near the threshold for a prediction of direct interaction given by the network extraction method. However, since it requires extensive noise for this additional edge to be predicted with even small frequency, and the resulting impact is, in any case, small, we have concluded a general robustness of the network extraction method. In all other networks recovered from the artificially introduced noise, the network topology was the same as in the absence of noise, with both edge weights and IL-23 concentration predictions at the time of experimental measurement predicted to only have a small relative difference. This result is an encouraging indication for the viability of focused network extraction for similar datasets.

The full recovered cytokine network model of the *in vitro* system as given in [98] identifies predictions for the regulatory cytokines of the system, IL-1α, IL-1β, IL-10, and IFNγ, from which the system can be fully described. We concentrated our further *in vitro* model analysis and the *in vivo* model extension on these regulatory cytokines and their interaction with IL-23. Using global sensitivity analysis, we determined that as expected, IL-23 concentration in the *in vitro* model was heavily influenced by the stimulatory effects given by LPS and IFNγ. The sensitivity to LPS and IFNγ is in line with the expected results as LPS is a powerful stimulator of immune response and IFNγ is a major regulator of immune response. With regards to the regulatory cytokines of interest, the sensitivity analysis of the *in vitro* system model indicated that the concentration of IL-23 depends most heavily on the direct down-regulatory effect of IL-10, identifying IL-10 as the most significant regulator of IL-23 concentration. IL-23 production was next most sensitive to the direct up-regulatory effects of IL-1β. The effect of IL-1α was small in comparison to the effects of other regulatory cytokines of interest in the system. To understand whether extracted cytokine networks from biopsies can inform our understanding more generally, we similarly conducted a global sensitivity analysis to the *in vivo* extension of the system model. We found that the introduction of clearance plays a dominant role in the *in*

vivo system model, substantially reducing the sensitivity of IL-23 production to the regulatory cytokines of interest in proportion to the clearance rate. While there was a decrease in sensitivity due to clearance, the ranking of sensitivity of the *in vivo* system model to the regulatory cytokines of interest was the same as that of the *in vitro* system for cytokines above a small threshold of sensitivity. Thus, IL-23 production is most influenced by the down-regulatory effects of IL-10, followed by sensitivity to up-regulatory effects of IL-1β and nearly no sensitivity to IL-1α effects in both models. With this preliminary insight given by our sensitivity analysis, we then proceeded to investigate the regulatory roles of each cytokine in detail. This included fluctuating the concentrations of cytokines to gauge responses of IL-23, as well as model reductions, equivalent to the full removal of a cytokine from a system or assay.

In both the *in vitro* and *in vivo* system models, there was an insensitivity to parameter values associated with IL-1α. This suggested a potentially simple model reduction, and we removed IL-1α from the model of both systems. In the *in vivo* system model, the removal of IL-1α resulted in essentially no changes to IL-23 concentration across clearance rates. However, in the *in vitro* model, undergoing model reduction of IL-1α created a relative difference of IL-23 concentration of over 10% at the experimental measurement time, and the relative difference also increased with time. This demonstrated a notable difference in the *in vitro* and *in vivo* models due to the impact of cytokine clearance, though this can be understood in that small differences in concentration would be amplified in time given one of the models did not have a steady state.

Next, we investigated the role of IL-1β in the regulation of IL-23 in both the *in vitro* and *in vivo* models. IL-23 is sensitive to IL-1β; however, the influence of IL-1β was initially unclear in both systems due to the indirect down-regulation and direct up-regulation of IL-23 by IL-1β, as seen in Figure 3.2. Here, we have demonstrated that the effect of IL-1β is always predicted to be up-regulatory in an *in vitro* setting. The impact of perturbing IL-1β was diminished as the concentration of IL-1β was increased, with progressively larger perturbations to IL-1β required to induce effective changes in IL-23 at $t = 16$ hrs. On the other hand, in the *in vivo* model, the changes to IL-1β concentration that we investigated resulted in essentially no change to IL-23 concentration at the experimental measurement time. However, while perturbations to the concentration of IL-1β are very small in the *in vivo* model, we may not remove IL-1β from the modelling equations completely without large relative changes to the steady state value of IL-23 (see Figures 3.12 and 3.13). Finally, as evident from the

network, and as expected, the effect of IL-10 is strictly down-regulatory, for both *in vitro* and *in vivo* models, and the effect of IFNγ is strictly up-regulatory in both system models.

While aspects of the predictions are essentially preserved compared to *in vitro* modelling studies, such as the sensitivity ranking of perturbations for IL-23 levels, other features of predicted cytokine responses differ on the introduction of *in vivo* clearance rates, highlighting the need for care in extrapolating predictions from cytokine biology from *in vitro* cultures to *in vivo* tissue biology. As we inspected the influence of individual cytokines with the introduction of clearance into the system, we observed differences in the impact of the regulatory cytokines from the *in vitro* system to the *in vivo* system models. Particularly, there were changes in the influence of IL-1α and especially IL-1β. These differences in behaviour of the *in vitro* and *in vivo* system models indicate that experimental studies based on very long clearance half-lives need not be fully representative of corresponding dynamics *in vivo* with short half clearance rates. In turn, this suggests that either experimental designs with cytokine clearance, or an additional modelling step including clearance, could be informative for the translation of insight from *in vitro* data to *in vivo* scenarios. Finally, given that the *in vivo* system model is highly stable, with a lack of bifurcations and excitability, alternative dynamics that may potentially arise in the *in vivo* setting may not be captured in initial experiment. However, the preservation of the relative ranking of sensitivity also indicates that *in vitro* studies are nonetheless informative, especially as such observations consider large regions of parameter space rather than the focused parameter estimates considered in Figure 3.12 for example.

In summary, we have validated the robustness of the network extraction method given in [98] to noise in the underlying data and described the behaviour of the system both *in vitro* and with the introduction of *in vivo* clearance, highlighting similarities and differences between the two models. We also have identified the dynamics predicted for the cytokine interaction network and analysed the impact of cytokines of interest on IL-23. We have found in the modelling framework that IL-23 is sensitive to the regulatory cytokine IL-10, a significant down-regulator of IL-23. We have also found that IL-1β is predicted to up-regulate IL-23 secretion in the *in vitro* system; however, these results are not observed in the extension of the current model to incorporate *in vivo* clearance, an exemplar result generating caution when extrapolating *in vitro* dynamics to the *in vivo* setting for the cytokine system. However, the conclusion that *in vitro* observations are unreliable for *in vivo*

inference is tempered by the observation that the relative ranking of sensitivity is preserved upon including *in vivo* motivated clearance rates. Further work is required to ascertain whether such observations apply to cytokine dynamics more generally. Furthermore, perturbation studies, such as the biopsy culture experiments utilised in the given model, are becoming more advanced with a combinatorial explosion of possibilities between cytokine interactions, motivating additional consideration of the nature and scalability of network extraction, as well as what larger data sets may reveal about the behaviour of larger cytokine interaction networks.

3.5 Conclusions and Future Directions

Our study contributes to the literature regarding cytokine network interactions, showing how indirect regulation pathways and network topology require further dynamical investigation in order to draw conclusions regarding target cytokine regulation. We further highlight differences in *in vitro* and *in vivo* model predictions, emphasising caution to mathematical modellers attempting to draw conclusions that translate across these settings. We note that in spite of the differences in *in vitro* and *in vivo* model predictions we detail, the relative sensitivity ranking is preserved between the predictions, underscoring the value of qualitative inferences derived from *in vitro* biopsy data.

Our findings on indirect interactions have significant implications in the control of downstream regulation of cytokines. This is of particular importance, as cytokine therapeutic treatments that target the regulation of specific cytokines often lose effectiveness over time. For example, current cytokine therapies treatment for IBD predominantly targets the cytokine pro-inflammatory cytokine TNFα, leading to remission of IBD in some patients [88, 47, 160, 48]. However, a proportion of patients treated with anti-TNFα therapies, ranging from 10 to 40% , do not respond or acquire resistance to TNFα inhibition therapies [189, 79]. Hence, the control of cytokines via alternative routes and upstream regulators, as investigated in our study, is of particular interest. Among our results, those regarding model reductions, which are comparable to the complete lack of production of a cytokine as well as its insensitivity to regulation, give insight into how biological mechanisms are affected by total or near total blocks, even if the target molecule is not directly interacting with the blocked cytokine.

Our insights into cytokine network topology influence may also be influential for experimental design in studies that apply network extraction and mathematical mod-

elling. That is, our results emphasise that the inclusion or exclusion of a cytokine with out-edges in the network is essential to its dynamics. This also foreshadows the increasing complexity of the network dynamics of the cytokine interaction network as the network grows. This suggests that for sparse data inference, individual perturbations of cytokines within experimental data generation may be required in order to accurately capture network edges. Finally, our sensitivity analysis compliments the body of literature indicating IL-10 and IL-1β as major regulators of IL-23 and contributes a description of their possible utility in the regulation of IL-23 in cytokine therapeutic treatments.

Future work regarding cytokine modelling of this form is highly dependent on the availability of experimental data. For example, the capture of characteristics of cytokine dynamics, such as excitability, necessitate the introduction of further measurement times. Additional time-points are also essential in the further modelling of *in vivo* cytokine dynamics. The introduction of high-throughput data, as well as the addition of time-point measurements, and the inclusion of more cytokines will lead to further insights and more opportunities for the understanding of cytokine network dynamics through mathematical modelling and simulation.

Part II

Temporal and Spatial Organisation of Mobile Traffic Response to Catastrophe

Chapter 4

Overview

4.1 Overview

The use of mobile phones and their corresponding applications is ubiquitous. As such, they provide a rich source of data regarding human behaviour and interests, mobility and transit patterns, and population dynamics. Insights from mobile phone usage data are varied and broad. For example, companies may use it to personalise services and advertising, improve operations and algorithms, and track user habits and tendencies to format their platforms for further engagement [11, 7, 179, 220].

According to a 2013 study, 91% of the 100 senior executives interviewed of Fortune 1000 companies were investing in big data projects [107, 7]. A 2024 survey across a variety of market sections including finance, healthcare, retail, and media stated that 87.9% of respondents claimed that investments in Data and Analytics are a top organisational priority [227]. Analysis of big data accounts for much more engagement, with industry giants such as Amazon generating 30% of its sales through analytics [7, 218].

Companies gather data and track user application behaviour to promote engagement on their platforms, integrating functionality from other popular applications. This can be seen in the integration of "Stories", a short lived social media image format popularised first by the application Snapchat in 2013 and later adopted by applications such as Instagram in 2016 and Facebook in 2017 [136]. Another example of this is the rise of short-form video content which lead to the popularisation of the application Tik Tok in 2019 and early 2020 leading to the corresponding launch of Intagram's short-form video feature "Reels" in late 2020 [175, 164, 34, 149]. More recently, the widespread popularity of artificial intelligence has lead to the integration of AI tools into the search platform Google and the social media platform X (formally known as Twitter) [110, 156]. These examples display the clear interest of companies

in trends of feature and functionality engagement in order to optimise their platform usage.

Analysis of mobile phone data patterns are also informative in epidemiology and the management of the spread of infectious disease. Mobile phone data can be used to track population movements and can provide key insight into contact patterns. Mobile phone data has been used both in modelling to help inform strategies regarding infectious disease outbreaks and has been applied to key interventions during epidemic spreading such as contact tracing and quarantining instructions during the COVID-19 pandemic [165, 22]. A prominent example of the use of mobile phone application data to manage epidemic spread was the NHS COVID-19 app, a part of the Department of Health and Social Care's NHS Test and Trace Programme in the National Health Service of the UK. The application alerted users whose phones had been in close proximity to other users who had tested positive for COVID-19, tracking and promoting the isolation of new cases of COVID-19 to prevent disease spread [165]. Mobile phone location patterns extracted from text and call data have been used to track the spatial spread of cholera [22], while search engine queries and social media have been used to track the spread and surges of influenza [2, 68]. These insights provided by mobile phone data unveil valuable contact and mobility patterns and give insight into human behaviour.

Among the many applications of mobile phone data, and that which we will be emphasising in this chapter, is its capacity to track the dissemination of information pertaining to catastrophic events. Data regarding mobile phone usage has been used to monitor and track disaster progression as well as crowd source information that can inform critical interventions such as emergency response, public warning systems, and coordinated evacuations. Mobile phone usage is critical to the dissemination of information to the general public. Additionally, given the ubiquity of mobile phones, any user may update working knowledge of a catastrophic event with many users circulating images, videos, and messages in the aftermath of catastrophic events. Mobile phone application usage during a catastrophic event further involves individual user interface with platforms such as Facebook rolling out a 'Safety Check' feature [150] that allow users to 'check in' as safe during a catastrophic event alerting other users and first-responders of the state of the catastrophe. This feature has been activated for a variety of catastrophic events including an earthquake in Nepal in 2015 [247], the Pulse Nightclub Shooting in the USA in 2016 [200], and a bombing in Brussels in 2016 [102]. The use of mobile phones extends to government agencies as well, with emergency alert mobile broadcast system examples including Japan's

tsunami warning system, the European flood warning system, and South Korea's missile test alerts. Applications which extend to mobile devices have also been used in scenarios that pertain to prediction displacement patterns and the improvement of humanitarian resource allocation, such as previous work using Facebook to estimate of the displacement of the Venezuelan population fleeing economic crisis and compare to government estimates [173].

Previous work using mobile phone data to detect anomalous events and understand response to catastrophic events has utilised call and GPS location data [240], phone-to-phone messaging [5], social media [93], network traffic data [13], and SIM data [23]. These are been used to make suggestions to network improvement given surges in traffic surrounding event occurrence [13] as well as prepare for and observe disaster [240] and analyse human response in its aftermath [23]. Patterns of human attention to disaster and shocking events using online platforms have also been extracted using Wikipedia and X [89, 170, 76].

Given these diverse applications of mobile phone data, our aims in this part of the thesis are multifold. In the following chapter, we will detail the use of mobile phone application data to track information spread in response to catastrophic events and the use of mobile phone applications for information dissemination through an exploration of these information patterns. We will elucidate the influence of different types of application platforms such as discussion-based or live-video based to disseminate and maintain engagement over time. Our analysis gives insight into how application features are used to shape information spread. Additionally, through our analysis of mobile phone application data, we examine human behaviour patterns regarding responses to catastrophe, offering a perspective into how people interact with the surrounding world using modern technologies.

In Chapter 5, we present our study of the detection and analysis of anomalous data patterns in response to catastrophe. We focus on categorising patterns of response in both space and time, tracking their relaxation over time. To this end, we use the NetMob2023 Data Challenge dataset, which provides mobile phone applications traffic volume data for several cities in France at a spatial resolution of a 100 x 100 m^2 square tiling and a time resolution of 15 minutes for a time period ranging from March to May 2019. We derive our analysis from mobile phone application traffic surrounding the fire of the Cathedral of Notre-Dame in Paris in 2019. We analyse the spread of information before, during, and after the catastrophic Notre-Dame fire on April 15th using volume of data uploaded and downloaded to different mobile applications as a proxy of information transfer dynamics. We identify different clusters of information

transfer dynamics in response to the Notre-Dame fire within the city of Paris as well as in other major French cities. We find a clear pattern of significantly above-baseline usage of the application Twitter (currently known as X) in Paris that radially spreads from the area surrounding the Notre-Dame cathedral to the rest of the city. We extend our analysis to a bombing that took place in the city centre of Lyon on May 24th. We detect a similar radial pattern in the city of Lyon in response to the bombing. Further, we present a null model of radial information spread and develop methods of tracking radial patterns over time. Overall, we illustrate novel analytical methods we devise, showing how they enable a new perspective on mobile phone user response to unplanned catastrophic events and giving insight into how information spreads during a catastrophe in both time and space. Methodological concepts used to develop this chapter and our analysis of the mobile phone application dataset are outlined in Sections 1.4.1 and 1.4.2, which introduce Kullback-Leibler Divergence and Clustering Analysis, standard methods and computational techniques used in data analysis through which our analysis will be interpreted.

Chapter 5

Detection of Anomalous Spatio-temporal Patterns of App Traffic in Response to Catastrophic Events

5.1 Introduction

Understanding how information propagates online during and after catastrophic events is an active field of investigation [75, 236, 125, 170, 89], with particular attention paid to patterns of data usage. Mobile phone data provides deep insight into the intricacies of human behaviour, with extremely high granularity on both temporal and spatial scales. Mobile phone and application datasets therefore enable large-scale data driven analysis applied to a wide range of areas including social network analysis [59], population dynamics [54], urban structure [134], public health [168], and disaster response [135]. Social media and online resources have been used to categorise and track the duration and intensity of responses to events and breaking news stories [125, 170, 76, 89, 133]. Attention has also been paid to the behavioural patterns that characterise and distinguish between planned and unplanned events [111, 50]. While most studies on social media or app usage have focused on single platform online activity [125, 111, 50, 242], recent work has emphasised the importance of cross-platform online data analysis to account for platform-specific effects [202, 8], such as platform design, features, and algorithms which may affect user engagement.

In this chapter, we analyse mobile phone data to understand how the temporal and spatial usage of different mobile phone applications are perturbed in the aftermath of an unprecedented event, and how these usage patterns evolve over time. We will focus on how a variety of mobile phone applications, which may have similar user-function,

respond to the same event. To this end, we use the NetMob2023 Data Challenge dataset [144] which provides mobile phone usage data for several cities in France for a range of applications over a period of time ranging from March to May of 2019 at a spatial resolution of a 100 x 100m^2 square tiling and a time resolution of 15 minutes. We investigate the dynamics and patterns arising from catastrophic events, most notably, one of the most extreme events occurring during this period: the fire at the Notre-Dame cathedral in Paris and the collapse of its historic spire.

The Notre-Dame Cathedral, built during the 12th and 13th centuries, has long been recognised as an emblem of French society. Located in the centre of Paris, it is a highly popular tourist attraction, as well as an important religious site. On April 15th, 2019, the roof of the Notre-Dame cathedral caught on fire and was severely burned. The fire broke out in the attic of the cathedral at 18:18 [176]. The occupants of the building, including tourists and worshippers taking mass, evacuated minutes later, at 18:20 [91], when a fire alarm sounded. Smoke was first visible by 18:43 from Paris' Left Bank [91], with firefighters officially called to the cathedral at 18:51 [24] when cathedral workers discovered the fire. The historic spire of the cathedral collapsed at 19:50 [176]. Given that the fire occurred at one of the most historic buildings in Europe, news spread quickly through both social media and traditional news media sources.

We analyse the spread of information before, during, and after the Notre-Dame fire using the volume of data uploaded and downloaded to different mobile applications as a proxy for information transfer. The methods we develop can be extended to other contexts to characterise mobile phone user response to unplanned catastrophic events, giving insight into information spread online during a catastrophe and the patterns of collective user attention response. We extend our spatial analysis to find consistent results in a second catastrophic event, a bombing in Lyon, France.

We present two kinds of analysis. First, we spatially aggregate data within each city and study the time-series of app traffic per city for a select subset of the available apps. We then use a simple anomaly detection method to determine whether traffic volume for each app spiked after the Notre-Dame fire, when those spikes occurred, and for how long. We use information about these spikes in traffic volume to cluster apps based on how they were used during the fire, attempting to discern patterns in type of application use both within the city of incidence of catastrophe and across other cities. We conclude that applications have a similar general user trend across cities, however there are variations in app traffic behaviour that can be explained by

the strong effect of city-specific user preferences. Additionally, we find that a fine-grained description of application function is a more accurate predictor of clustering behaviour, emphasising the role of platform features in user-engagement. Second, we analyse how traffic spikes were distributed spatially throughout the city of Paris over time as news of the fire spread and mobile phone users sought more information through social media, apps, and streaming services. Most prior studies of information spread focus on the relationships between users in social-network space [161, 85, 42]. There are few works, however, which investigate the geo-spatial dependence of the spread of information over social media [243]. We find that that the abnormally elevated traffic appears to spread radially from the centre of the catastrophe. We construct a null model to demonstrate how everyday data patterns shift to radial data patterns in response to catastrophe followed by a subsequent relaxation of radial patterns over time. This is consistent with other literature [14], as well as our own extended analysis to a bombing occurring in Lyon, France. We propose methods for quantifying this radial pattern of the information spread, noting that this pattern of information spread also extends to planned events. A graphical introduction to our analysis is given in Figure 5.1.

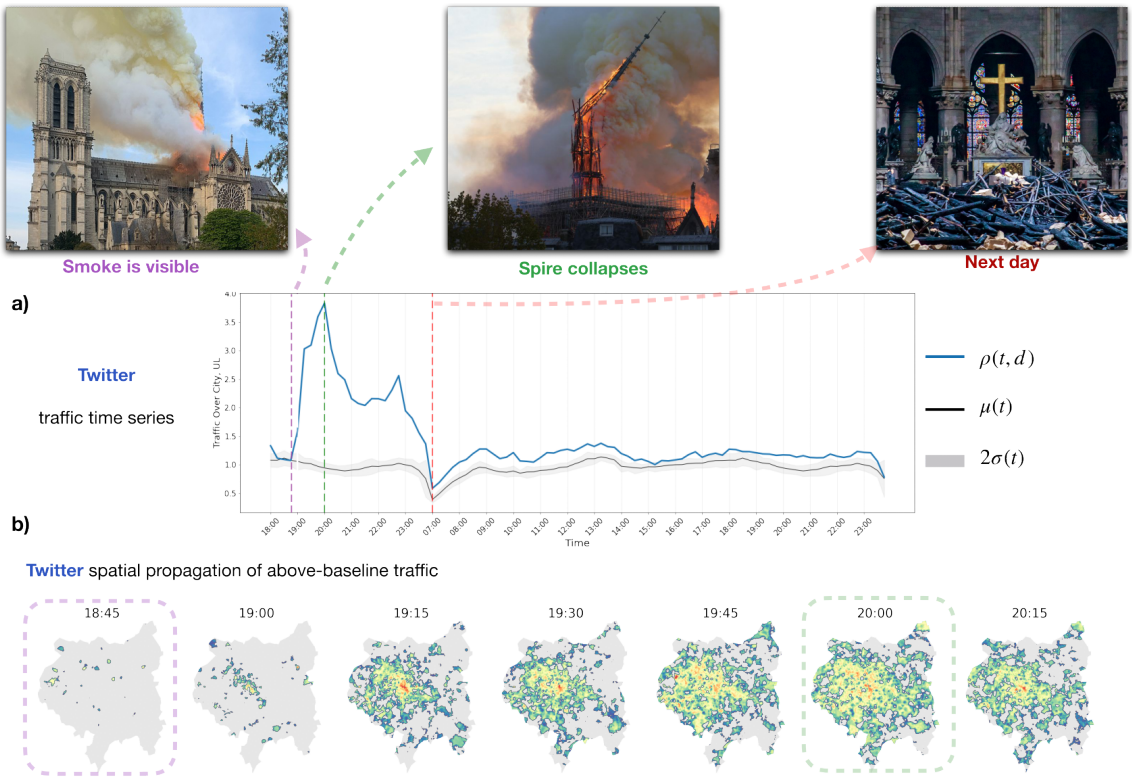

Figure 5.1: **Graphical introduction: a)** The time series evolution of the application Twitter from 18 : 00 to end of day on the day of the fire and 7 : 00 to end of day the day following the fire. The traffic on the day of the fire, ρ, is given in blue, and the mean (μ) and two standard deviations (2σ) of traffic in previous weeks is in grey. **b)** The spatial evolution of application traffic of Twitter is shown in time on the outline of the city of Paris, with intensity of app traffic from an established baseline of uplink traffic displayed.

5.2 Methods

In this section, we will describe the methodology pertaining to the detection and analysis of patterns of spikes in application traffic volume both temporally and spatially.

5.2.1 Data

The dataset under scrutiny is the NetMob2023 Data Challenge dataset [144], which provides the traffic volume for 68 mobile applications in 20 cities in France at a spatial resolution of a 100 x 100 m^2 square tiling and a time resolution of 15 minutes for 77 days, ranging from March 16th to May 31st 2019. The full list of applications available in Appendix B: Part II. The NetMob2023 dataset contains both uplink (transmissions

to the cell tower) and downlink (transmissions from the cell tower) application traffic volume. We consider both of these traffic directions simultaneously by summing the two. We limit our analysis to data from Paris, Marseille, Lyon, Montpellier, Rennes, and Strasbourg. We note that the dataset does not contain information regarding content transferred along the cellular network.

In our analysis, we consider a subset of applications in the dataset after sorting for relevance (e.g. exclusion of mobile phone games) and ability to distinguish app traffic as being in some form related to the fire (e.g. exclusion of general internet browsing such as Yahoo). The exclusion of general internet browsing applications is due to the inability to determine which feature or website is being used by the browsing site, making this uninformative to our aims. Our criteria for included apps consist of applications for social media, messaging, transportation and location, video-sharing and streaming, and cloud storage. We note that cloud storage is included, as spikes in traffic are indicative of photo and video interaction. That is, cloud storage applications traffic is generated by photo and video viewing, downloading, and storage. We recover the following list of 24 applications that will be assessed for traffic spikes:

- Apple App Store
- Apple Video
- Apple iCloud
- Apple iMessage
- DailyMotion
- Facebook
- Facebook Live
- Facebook Messenger
- Google Drive
- Google Maps
- Google Play Store
- Instagram
- Molotov
- Orange TV
- Periscope
- Snapchat
- Telegram
- Twitch
- Twitter
- Uber
- Waze
- Web Transportation
- WhatsApp
- YouTube

Of the applications considered, 11 of 24 applications experience traffic spike. The full table of all spiking app is found in Table 5.1 in Section 5.3. None of the transportation or location applications considered experience abnormally high traffic spikes as defined by our methodology. Other applications that do not experience traffic spikes, contrary to initial expectations, include the messaging app Telegram and the video streaming app YouTube. Excluded applications were not analysed for potential traffic spikes.

5.2.2 Detecting Spikes in App Traffic Volume

In this dataset, each city c is represented spatially by a tiling (or tessellation); we refer to the tiles of a city c as a list $w_c = 1, \ldots, n_c$, where n_c is an integer corresponding to the total number of tiles in that city. We then denote the volume of traffic for app α in city c at tile $i \in w_c$ at time t on day d as $\rho_{\alpha,c,i}(t, d)$.

In order to form a temporal description of the traffic volume, we first perform a spatial aggregation of the traffic for each app over all tiles of the city within a day d,

$$\rho_{\alpha,c}(t, d) = \sum_{i \in w_c} \rho_{\alpha,c,i}(t, d), \tag{5.1}$$

to recover a time series representing the total traffic in the city at each 15 minutes time interval t for the day d in the dataset.

We utilise these time series to calculate baseline statistics against which to measure abnormal app traffic. Since the fire of the Notre-Dame Cathedral occurred on Monday April 15th, we construct the baseline and standard deviation based on app traffic from the weekdays, Monday through Thursday, preceding the fire in the dataset: March 18th to April 11th 2019. We exclude Fridays and weekends as we expect they will have more heterogeneous application traffic patterns. We also confirmed that no national holidays coincided with our baseline period. We denote the mean traffic volume over this period for application α and city c as

$$\mu(\alpha, c, t) = \langle \rho_{\alpha,c}(t, d) \rangle_d, \tag{5.2}$$

where the average $\langle \cdot \rangle_d$ is taken over all weekdays d in the baseline period. The value of $\mu(\alpha, c, t)$ at each time t corresponds to the average traffic volume for the app α at each measured 15 minute time-point t. We also compute the standard deviation for each tile and time and denote it $\sigma(\alpha, c, t)$. For convenience, from here on we will drop the indices for app and city, and simply use $\mu(t)$ and $\sigma(t)$ to represent the average and standard deviation for any combination of app and city over the baseline period.

Similarly, we will use $\rho(t, d)$ to refer to a time series on a specific day.

We now define the *distance from baseline* at time t on day d as the following modified z-score,

$$D(t, d) = \frac{\rho(t, d) - (\mu(t) + 2\sigma(t))}{\mu(t) + 2\sigma(t)}, \tag{5.3}$$

indicating the value of traffic volume in relation to two standard deviations from the baseline mean of traffic. We will utilise the shorthand $D(t)$ under the assumption we have selected a day of interest.

Finally, we define a spike in the traffic of an app as a period of at least 3 consecutive time points $[s_{start}, s_{end}]$ where $D(t) > 0$, in order to account for noise. For symmetry, we define the end of a spike as a period of three consecutive time points $[s_{end}+1, s_{end}+3]$ where $D(t) < 0$.

5.2.3 Temporal Clustering

Traffic volume spikes of applications exhibit different characteristics in terms of duration, amplitude, and position in time. In this section, we define a set of features on the $D(t)$ time series of an application and use these features as input dimensions in the k-means clustering algorithm to characterise apps with similar traffic spiking patterns. We will present two analysis of the temporal dataset. In the first, which we consider the main analysis of this work, we consider the analysis of the data in two days, the day of the fire and the day after the fire. The observation of two days allows us to consider the short- and long- term behaviours of applications separately. This allows us to consider clusters of applications in terms of immediate reactions to the fire while also identifying platform features and user behaviour which contribute to continued engagement on platforms. Our secondary analysis considers the joint two-day clusters of application time-series. This compliments the main analysis by providing a more holistic view of event responses, identifying clusters as shifts in patterns over time instead of focusing on the more detailed reaction patterns. In both analysis, we introduce an overnight gap on the day after the fire from midnight to 7 : 00, since baseline week nights have small variation compared to weekdays, meaning smaller standard deviations and more noise in the use of our anomaly detection method. This gap corresponds to data in this period not being considered.

Considering that the smoke during the fire of Notre-Dame is first visible 18:43 on April 15th, we denote a spike as occurring from 18 : 45 of April 15th to the end of that day as spike 1, and a spike occurring from 7 : 00 to 24 : 00 of April 16th as spike

2. Our analysis computes at most one spike per day, with only a first possible spike being considered.

The 5 features that we compute are:

- (s_i^{start}, s_i^{end}): the starting and ending times of the ith spike
- $s_i^{duration}$: the duration of the spike
- s_i^{max}: the maximum value of $D(t)$ during the spike
- $s_i^{aggregate}$: The sum of all $D(t)$ values during the spike

This methodology is consistent across both the main and secondary analysis. In the main analysis the traffic time series of each app are represented as two features vectors of length 5 with the features corresponding to the day of and the day after the fire, clustered within their respective days. In the secondary analysis, the traffic time series of each app are represented by a single features vectors of length 10. This 10 dimensional vector is composed of the union of the 5 dimensional features vectors for an application across the day of and day after the fire. We normalise the features vectors by the maximum and minimum values for spiking apps within each feature for the day. Finally, we perform a k-means clustering of the feature vectors. Given the difference between the number of apps that spike in Paris compared to the other cities considered, as well as our interest in understanding if applications will behave differently in the city of incidence of catastrophe, we cluster apps in Paris separately from all other cities. Thus, in the main analysis we compute clustering for the following groupings independently: Paris on April 15th, Paris on April 16th, other cities on April 15th, and other cities on April 16th. We choose number of clusters $k = 5$ by computing the within-cluster sum of squares errors over values of k and using the elbow method to manually inspect the curves. In the joint two-day cluster analysis, we also compute clustering for Paris and other cities independently. Following the same methodology, we recover $k = 6$ clusters for Paris and $k = 5$ clusters for other cities. Features values and the curves used to determine the optimal number of clusters are given in Section 5.3.

5.2.4 Spatial Analysis

In this section, we describe the methods used to investigate the spatial patterns of the increase in anomalous app traffic in response to the fire, which will then be extended to other events.

5.2.4.1 Traffic Spikes in Space

Our spatial analysis is primarily focused on to the traffic response to the fire of Notre-Dame for the application Twitter in the city of Paris. Paris is considered as this is the city of incidence of the fire, allowing for measurement with respect to the epicentre of catastrophe. Twitter is considered as this application experiences both a strong and long-term response, giving the possibility of analysing its behaviour over time. We later extend the spatial methodology developed to Whatsapp and Periscope during the Notre-Dame fire. Finally, we apply our methods to a bombing in Lyon which indicates similar radial patterns and conclude by showing the our event detection methodology is also applicable to other events.

We first recall that the Netmob2023 Data Challenge dataset has a spatial resolution of 100 x 100 m^2 given as a square tiling of the city of interest. Hence, all spatial analysis can be conducted at this spatial resolution. Recalling the definition, in Equation 5.1, of the volume of traffic, we will consider the time series of $\rho_{\alpha,c,i}(t,d)$ for the city of in which we will conduct the analysis (ex. c =Paris), and for the specified application(ex. α =Twitter). Now for *each tile* $i \in w_c$ we define the tile-wise time-series as $\rho_i(t,d)$.

We use these tile-wise time series to calculate a baseline mean and standard deviation for the traffic for each spatial tile. Analogously to what is presented in the previous sections, we construct our baseline statistics based on application traffic on the weekdays Monday through Thursday, again excluding Fridays and weekends as we expect they will have different patterns. For our Twitter analysis in Paris following the fire of Notre-Dame this is from March 18th to April 11th excluding Fridays and weekends. For our analysis of the bombing in Lyon, given that this bombing occurs on Friday, May 24th of 2019 at 17 : 28 [86], we construct baseline statistic from all Friday's prior to the bombing within the dataset. We denote the mean traffic volume over the baseline statistic period for a tile i similarly to the above section as,

$$\mu_i(t) = \langle \rho_i(t,d) \rangle_d. \tag{5.4}$$

The value of $\mu_i(t)$ at each time t now corresponds to the average traffic volume for the selected application at the 15 minute data time-point t in tile i of the spatial grid of the selected city. We also compute the standard deviation from our baseline statistics for each tile and time and denote it $\sigma_i(t)$.

We can now define the tile-wise *distance from baseline* at time t on day d, in tile i as the following modified z-score:

$$D_i(t,d) = \frac{\rho_i(t,d) - (\mu_i(t) + 2\sigma_i(t))}{\mu(t) + 2\sigma(t)}. \tag{5.5}$$

We can then compute $D_i(t)$ for the app on each tile i of the city. In our analysis of the fire, this is begins at $t = 18:45$ on April 15th, the day of the fire, until $t = 24:00$. In the analysis of the bombing this begins at $t = 17:30$ on the day of the bombing.

5.2.4.2 Quantifying Spatial Effects

We are interested in how information propagates across applications with respect to the source of the catastrophic event. That is: what spatial patterns emerge in response to catastrophic events, how can we quantify these patterns, and how do they evolve temporally? Since the concept of physical distance does not exist in the virtual application space, it is of particular interest to understand how quickly information propagates across the internet and how this translates to analogous physical spatial patterns.

In order to analyse physical spatial patterns we take inspiration from Collective Response of Human Populations to Large-Scale Emergencies by Bagrow et. al [14], which considers anomalous traffic volume as it pertains to the distance from the epicentre of the catastrophe. The analysis presented in [14] utilises call activity to examine if event information is constrained to the area surrounding a catastrophe or if there is a wider spatial distribution of anomalous activity. In the analysis, call volume is integrated over time periods of 2 hours. Then concentric circular shells of radius r over this integrated call volume centred on the epicentre of the catastrophe are examined over a variety of disaster events. Bagrow concludes that anomalous activity experiences a peak at the epicentre of the catastrophe, with an approximated exponential radial decay outward from the epicentre.

Following this, we begin our analysis by constructing concentric square shells of increasing radial size centred on the epicentre, with radius being the diagonal of each square and the distance from the epicentre measured in kilometres. A shell is defined as the collection of all of the tiles included in the square of the current radius and excluding all of the tiles from the previous radial distance, in essence a square annulus. For each shell, we sum the total anomalous traffic, $\sum_{i \in r} D_i(t)$, within the shell and normalise by the number of tiles within the specified shell resulting in a measure that

shows the mean tile-wise volume of application traffic as a function of the radius [14],

$$\Delta D(r) = \frac{\sum_{i \in r_m} D_i(t)}{\sum_{r_m} i}, \forall m = 0, 1, 2, ...M. \tag{5.6}$$

We define radius zero to be the single tile at the epicentre of the event. We consider a maximum shell radius to be the value of the last tile before exiting the defined city tiling in any direction. This consideration is sufficient in our analysis to recognise spatial patterns, however, this directional consideration may be adjusted, particularly for events which may occur on the edge of a defined city tiling where radial shells will quickly exit the city in a certain direction.

At each radius, we now have a measure of abnormal spatial traffic which allows us to quantify changes in traffic at all radial distances. We posit that given Equation 5.6, radial outward spread from the epicentre would be represented with a decaying function, with the rate of decay pertaining to the intensity of the decay of anomalous application traffic activity.

The authors in [14] generally consider change in anomalous traffic volume from the epicentre aggregated over some time period, giving the course-grained, general behaviour of the response to catastrophe. However, since we are interested in how the radial spread of information relaxes, we consider the 'instantaneous' function of Eq. 5.6 at each measured time point. This more fine-grained approach to understanding radial spread shows how radial patterns evolve over time and relax from perturbation. Additionally, looking at individual snapshots of anomalous traffic volume ensures that the response of one measured time does not dominate an interval of measurement.

Now that we have a measure of radial spread that can be analysed at each time point, we can examine how radial the spread of information is and how quickly the abnormal application traffic patterns relax. To do so, we begin by using linear regression to recover a best-fit slope for the data at each time-point, tracking how the slopes of the recovered fits change throughout the course of the event. The radial decay is stronger as the value of the slope decreases.

We further explore the comparison of the data distribution before the event and its change to a radial spread following the event by using Kullback-Leibler (KL) divergence. We first calculate the aggregated function of Eq. 5.6 two hours after the event, constructing a course-grained reference distribution of the activity pattern in the time following the event. We then measure the KL divergence of each instantaneous snapshot in the form of Eq. 5.6 to this aggregated function. This allows us to quantify the temporal evolution of the spatial distribution of anomalous application

traffic, highlighting differences in behaviour before, during, and after the fire relative to the reference distribution immediately following the catastrophe.

We then extend our analysis by constructing a null model of optimal possible radial spread given the data. The null model is constructed for a time point by considering all values of abnormal traffic for spatial tiles in the maximum radius searched in that time point. These values are then sorted by maximum value outward from the epicentre. That is, the maximum value of all tiles in the snapshot is assigned to the epicentre of the event, the next n highest values are then assigned to the n tiles within the first radius considered away from the epicentre, and so forth. We can then calculate a function of the form of Eq. 5.6 on this optimal radial spread according to the data. As a next step, we compare the true data distribution to the optimally radial null model at each point in time by measuring the KL divergence between the two distributions. Finally, we plot the measured KL divergences over time between the instantaneous snapshots to both the aggregate and null models, tracking how the spread of information approximates a radial spreading pattern and relaxes from the radial state over time.

5.3 Results

This work captures how the usage, or *volume of traffic*, of different applications varies in response to a catastrophic event. We will investigate what patterns, both temporally and later spatially, of application traffic response emerge. From these insights, we will draw conclusions about user-preference and behaviour in response to catastrophic events.

5.3.1 Temporal Analysis: Anomalous Application Traffic in Response to the Fire

We begin by identifying applications whose time series of traffic volume present an above-baseline increase in response to the fire. We consider the daily baseline of traffic volume for each app, for each 15 minutes interval of a day of the week (Monday to Thursday, from 00:00 until 23:45), as the average value of traffic volume computed over the same 15-minutes interval on all weekdays. We then define a measure of *distance from baseline* $D(t, d)$, at time t on day d. We thus consider a spike in the traffic of an app a time-span of at least 3 consecutive time points where $D(t, d) > 0$. Recall that we consider Paris and other cities separately to understand whether other cities

react differently than the city of incidence and if users have city-specific application behaviour patterns.

In Table 5.1, we list the 11 apps that present spikes of traffic in response to the fire in the city of Paris; we also identify subsets of the same apps whose traffic volume spikes above baseline in the cities of Lyon, Marseille, Montpellier, Rennes and Strasbourg. Abnormal app traffic volume spikes, both in Paris and in other French cities, are consistent with the assumption that human activity spreads across social-media as people react in real-time to catastrophic events. Interestingly, the applications spiking in cities other than Paris are always a subset of the applications that spike in the French capital, the city of incidence.

Application	Pa	Mrs	Ly	Mtp	Rn	Sg
Apple Video	✓		✓	✓		✓
Apple iCloud	✓					
DailyMotion	✓			✓		✓
Facebook	✓			✓		✓
Facebook Live	✓					
Facebook Messenger	✓	✓	✓	✓	✓	✓
Instagram	✓		✓	✓		
Molotov	✓		✓	✓		✓
Periscope	✓	✓	✓	✓	✓	✓
Twitter	✓	✓	✓	✓	✓	✓
WhatsApp	✓			✓		

Table 5.1: Comprehensive table of all apps that spike in Paris (Pa), Marseille (Mrs), Lyon (Ly), Montpellier (Mtp), Rennes (Rn) and Strasbourg (Sg).

We posit that the overall spike in app usage in Paris is reflective of the heightened interest in the city of incidence. Further, we posit that some apps spike in some cities and not others due to within-city application preference. this could be largely a demographic, such as age group, income, or lifestyle related or a city-cultural phenomenon, apps are popular in a city and users reflect what users around them use. Additionally, we suggest that certain apps spike in some cities but not other due to within-city preference. This may be driven by demographic factors, such as age group, income, or lifestyle related, or could be be a city-specific cultural phenomena, where an app's adoption in a city is influenced by users around them.

5.3.2 Two-Day Analysis: Short and Long Term Behavioural Clusters and Patterns

In the following sections we aim to identify groups, or clusters, of similar application traffic spiking patterns that correspond to various responses to the event or shed light on the ways in which the news of the fire is spread across platforms and how users engage with these platforms. This includes both the short and long term responses to the incident, *i.e.*, in the hours immediately following the fire of Notre-Dame on April 15th and on the following day. We consider the day of and the day after the fire separately, clustering immediate and extended responses as distinct properties of the applications.

5.3.2.1 Patterns on the Day of Fire

First, we define five features that describe the spiking patterns observed in the individual traffic volume time series for the initial post-event response:

- the starting time of the spike, s_1^{start};
- the end of the spike, s_1^{end};
- the maximum value of the spike, s_1^{max};
- the length of the spike, $s_1^{duration}$;
- the sum of all values $D(t, d)$ in $[s_1^{start}, s_1^{end}]$.

We then perform a k-means clustering (see Chapter 1 for further details) with $k = 5$ on the spiking-features vectors of all apps that spike within the city of Paris (Figure 5.2.a), and all the apps that spike in the other selected French cities (Figure 5.2.b). Section 5.3.3, Cluster Selection and Feature Values presents the elbow method used to determine k (see Figure 5.5) and reports the selected features of each application (see Table 5.6 and Table 5.7).

In Figure 5.2 each cluster consists of applications whose traffic time series display similar responses to the event as characterised by their features vector. We represent the typical response pattern of apps assigned to each cluster as a normalised radar plot of the *centroid* of each cluster (Figure 5.2.a, left and Figure 5.2.b, left). The radial axes of the radar plot correspond to the previously defined spiking features. The suffix 1 in s_1 refers to *first* day, or day of fire. We note that the spiking features are normalised with respect to its minimum and maximum values, across all apps,

found for the city of Paris (Figure 5.2.a), and for the other cities (Figure 5.2.b). Next to each radar plot, we show a representative time series of $D(t, d)$ of one of the applications assigned to the relative cluster (Figure 5.2.a, right and Figure 5.2.b, right). The applications for clusters in Paris and other cities are given in Table 5.2 and Table 5.3.

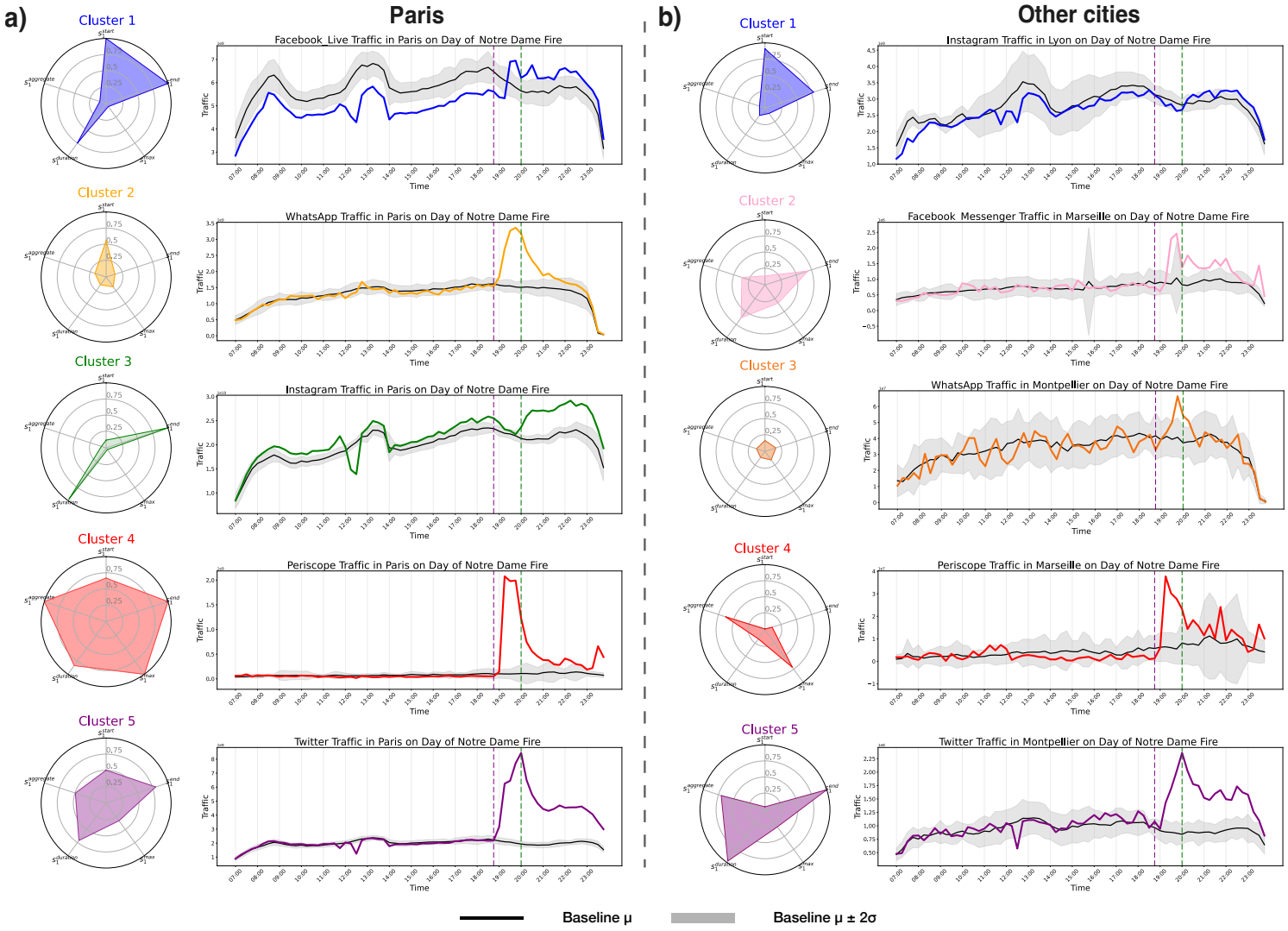

Figure 5.2: **Clusters and Representative Time Series on Day of Fire: a)** For the city of Paris, the radar plots of clusters of applications are given on the left, showing the value of different features assigned to each cluster. On the right of each radar plot is a time series for an application that is representative of the applications in that cluster. Note, the information in the radar plots for each cluster is normalised within the city of Paris. The applications in each cluster are given in Table 5.2. **b)** For all other cities being considered, the radar plots of clusters of applications are given on the left, with a representative time series of an of an application in the cluster shown on the right. Note, the information in the radar plots for each cluster is normalised for all cities excluding Paris. The applications in each cluster are given in Table 5.3.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
DailyMotion	Apple Video	Apple iCloud	Periscope	Twitter
Facebook Live	Molotov	Facebook		
	WhatsApp	Facebook Messenger		
		Instagram		

Table 5.2: Comprehensive table of all apps per cluster, in Paris on the day of fire, 15th April 2019. See Figure 5.2.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Instagram Ly	Facebook Messenger Mrs	Apple Video Sg	Periscope Mrs	DailyMotion Mtp
	Facebook Messenger Sg	Apple Video Mtp	Periscope Sg	Facebook Mtp
	Facebook Messenger Ly	DailyMotion Sg	Periscope Ly	Instagram Mtp
	Facebook Messenger Mtp	Facebook Sg	Periscope Mtp	Twitter Mrs
	Facebook Messenger Rn	Molotov Sg		Twitter Sg
		Molotov Ly		Twitter Ly
		Molotov Mtp		Twitter Mtp
		Periscope Rn		
		Twitter Rn		
		WhatsApp Mtp		

Table 5.3: Comprehensive table of all apps per cluster spiking on the day of fire, 15th April 2019, for all other cities: Marseille (Mrs), Lyon (Ly), Montpellier (Mtp), Rennes (Rn) and Strasbourg (Sg). See Figure 5.2.

In the city of Paris, we recover from our clustering methods three clusters of applications with similar behavioural patterns and two outlier clusters containing the singular behaviour of an application. The applications in each cluster are given in Table 5.2. Cluster 1 is characterised by low-amplitude, long spikes with late start times relative to other clusters (high s_1^{start}), yet sustained until the end of the day (high $s_1^{duration}$ and s_1^{end}). Apps in this cluster are thus characterised by a delayed and moderate, yet sustained response to the spreading of the news of the fire. Cluster 2 is composed of apps with a late spike of relatively short duration and moderate amplitude. The response to the fire is closer to the collapse of the spire, rather than the moment in which the fire becomes visible, yet the perturbation of the traffic time series of apps in this cluster is the shortest w.r.t. the other clusters. Applications in Cluster 3 are characterised by long, low-amplitude spikes, with very early start times (low s_1^{start}). Apps in this clusters, such as Instagram, see a rise above baseline of their traffic volume as soon as the fire becomes visible from the outside of the cathedral;

the traffic volume stays persistently above baseline throughout the rest of the day, yet the intensity of such anomalous activity is moderate.

The first outlier cluster, Cluster 4 includes one app with a distinctive behaviour, Periscope. Periscope is a live video streaming app: its traffic time series has a massive spike (10 times larger than average over the previous weeks in the same time points) at 19:00, *i.e.*, in the 15 minutes immediately following the smoke. The amplitude of the spike then decreases rapidly after 45 minutes, yet it stays above-baseline until the end of the day. This pattern of usage regarding Periscope is unique to the application, showing a strong preference in user interface. This cluster also characterises the behaviour of a live-streaming platform with users interacting shortly after the event in its immediate aftermath and then quickly losing interest in live footage over time.

Finally, the second outlier cluster, Cluster 5, contains only the discussion-based social media platform Twitter. The large, well sustained spike that characterises Twitter traffic in Paris is particular to the response to catastrophe on this platform.

In other cities, we also recover 5 clusters of applications, giving further insight into user interface preference within and across cities. The applications and their respective clusters are illustrated in Table 5.3. Notice that although the same application in different cities mostly belong to the same clusters, this is not always the case. We also obtain one outlier cluster which contains a single application, Instagram in Lyon, which demonstrates completely distinct behaviour from both other application usage in Lyon, and Instagram usage in other cities. This emphasises the conclusion that variations in application traffic behaviour is strongly related to city-specific user preferences, and usage dynamics.

The app in Cluster 1, Instagram in Lyon shows a similar, delayed spike as Cluster 1 in Paris, yet the duration of such above-baseline traffic is very short. Cluster 2 comprises apps characterised by early and sustained spikes, that however relax back to baseline well before the end of the day. Cluster 3 groups together apps whose traffic is perturbed early in response to the fire, however the amplitude and duration of such spikes are very low in comparison with apps in the other clusters. Cluster 4, which includes the application Periscope for almost all other cities, experiences a short but intense spike of large amplitude, coherently with what is observed in Paris. Finally, Cluster 5 is composed of applications experiencing elevated levels of application traffic that last until the end of the day.

While these patterns have slight variations between the 5 types of behaviour in Paris and the 5 types of behaviour in other cities, there are notable similarities in user usage for applications in response to catastrophe. This is particularly relevant

to the outlier clusters in the city of Paris and how they translate to the across-city clusters. Both in Paris and in other cities, the live video streaming app Periscope is characterised by a very large, yet brief, spike in usage, taking place shortly after the start of the fire. Like in Paris, a single cluster (Cluster 4 of other cities) is characterised only by application traffic patterns of the app Periscope across 4 other cities. This result shows the consistency of user interaction with this specific platform, showing that user tendencies can conform across an application with a narrow scope, such as live video streaming. Patterns that are similar in the French capital and in the other cities are also evident in the behaviour of Twitter. For 4 of 5 cities, Twitter clusters into Cluster 5. However the participation of the Twitter application in many cities in this cluster is shared with other apps, unlike in the case of Paris where Twitter is an outlier that does not cluster with other applications. This showcases the existence of general trends in app usage across cities as well as city-specific user preference effects. Applications with the behaviour pattern of the cluster primarily represented by Twitter undergo a large spike after the start of the fire, have a sharp increase in traffic until the collapse of the spire of the cathedral, and remain persistently well-above baseline during the rest of the day. This is true for both Paris and other cities. For both Paris and other cities there is also a cluster (Cluster 1 in both Figures 5.2.a and 5.2.b) with shallow spikes that start relatively late with respect to other applications.

We further consider whether applications in the other cities, excluding Paris, cluster similarly across these cities. We find that while this is mostly the case, this phenomenon is not exclusive. Applications that spike in multiple cities tend to cluster together, for example, Facebook Messenger clusters together for all the cities. However, there is some slight variation of application usage based on within-city user behaviour. This is evident in that there are apps that spike in some cities, and not others. For example, Instagram and Molotov spike in some cities, but not in all. Furthermore, there is some slight variation of clustering, such as how Periscope primarily clusters into Cluster 4, yet Periscope in Rennes is in Cluster 3. The same can be seen with Daily Motion, which appears for only 2 of the other 3 cities considered, in separate clusters. These results suggest the strong affect of city specific user-preference.

We find that applications with the same broad *'function'*, such a social media platforms, can exhibit very different response patterns. One may have better success in predicting app response patterns by a more fine-grained type such as, messaging,

broadcasting, image sharing, live video, etc. For example, the reaction-based platforms Facebook Live and DailyMotion cluster together, but don't cluster with other platforms that could generally be considered social media. This result emphasises the importance of application interface features in maintaining user attention.

5.3.2.2 Patterns on the Day after Fire

We now consider the long-term behaviour of applications and their relaxation from abnormal traffic both in Paris and in other cities by considering app traffic on the day after the fire. We begin our analysis at 7 : 00 on the following day to account for small baseline standard deviations during the night. In Figure 5.3 we show app clustering for Paris (Figure 5.3.a) and other cities (Figure 5.3.b) on the day after the fire of Notre-Dame. The applications and their respective clusters are illustrated in Table 5.4 and Table 5.5. Section 5.3.3, Cluster Selection and Feature Values shows the selection of k in the use of the k-means algorithm is shown in Figure 5.7 and the value of the features in the cluster is given in Table 5.8 and Table 5.9.

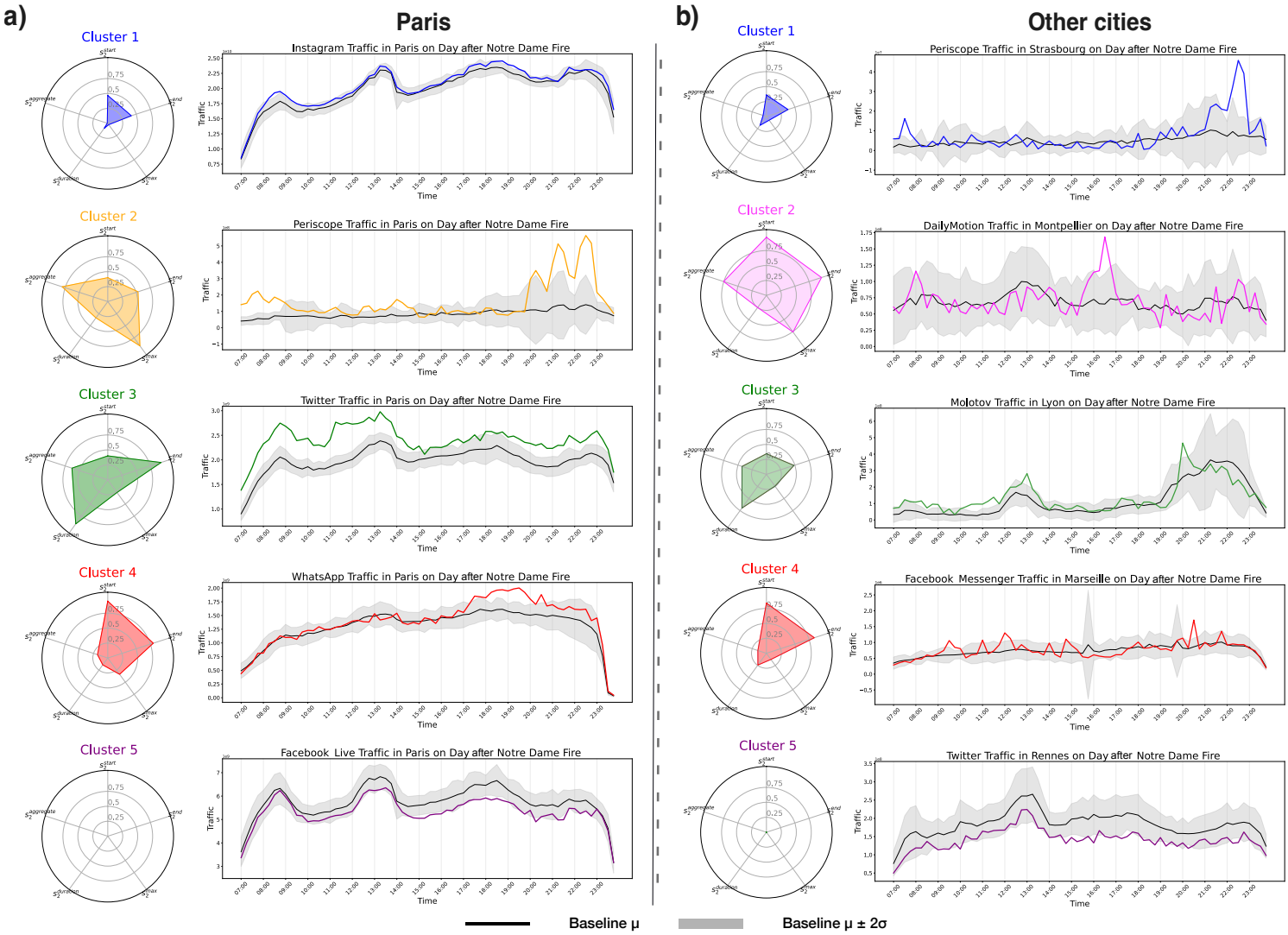

Figure 5.3: **Clusters and Representative Time Series on Day after Fire: a)** For the city of Paris, the radar plots of clusters of applications are given on the left, showing the value of different features assigned to each cluster. On the right of each radar plot is a time series for an application that is representative of the applications in that cluster. Note, the information in the radar plots for each cluster is normalised within the city of Paris. The applications in each cluster are given in Table 5.4. **b)** For all other cities being considered, the radar plots of clusters of applications are given on the left, with a representative time series of an of an application in the cluster shown on the right. Note, the information in the radar plots for each cluster is normalised for all cities excluding Paris. The applications in each cluster are given in Table 5.5

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Apple iCloud	Molotov	Apple Video	DailyMotion	Facebook Live
Facebook	Periscope	Twitter	WhatsApp	Facebook Messenger
Instagram				

Table 5.4: Comprehensive table of all apps per cluster, in Paris on the day after the fire, 16th April 2019. See Figure 5.3

.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Periscope Sg	DailyMotion Sg	Molotov Ly	Apple Video Ly	Apple Video Sg
Periscope Rn	DailyMotion Mtp	Molotov Mtp	Apple Video Mtp	Facebook Sg
	Facebook Mtp	Twitter Mtp	Facebook Messenger Mrs	Facebook Messenger Sg
	Periscope Mrs		Periscope Ly	Facebook Messenger Ly
	Periscope Mtp			Facebook Messenger Mtp
				Facebook Messenger Rn
				Instagram Ly
				Instagram Mtp
				Molotov Sg
				Twitter Mrs
				Twitter Sg
				Twitter Ly
				Twitter Rn
				WhatsApp Mtp

Table 5.5: Comprehensive table of all apps per cluster, on the day after the fire, 16th April 2019, in the other cities: Marseille (Mrs), Lyon (Ly), Montpellier (Mtp), Rennes (Rn) and Strasbourg (Sg). See Figure 5.3.

We note that for most applications, there is a quick general relaxation of spiking behaviour, that is, the abnormal app traffic does not last into the day after the fire. However, there are few exceptions, particularly in the city of Paris. The discussion based platform Twitter shown in Cluster 3 remains with elevated traffic levels throughout the second day, possibly due to new user content (*i.e.* discussion) being generated on the forum, unlike in the case of an image-based platform like Instagram. Notably, there is some spiking behaviour in clusters both for Paris and for other cities in the evening, as can be seen in Clusters 3, 4, and to a lesser extent, Cluster 1. This is particularly notable in the spiking of Periscope. We posit that this spiking, such as that of the video application Periscope and the messaging platform Whatsapp are associated to the organising of and sharing of content concerning a vigil held in response to the fire at the Cathedral of Notre-Dame, which occurred approximately 24

hours after the fire and lasted into the night [153]. Additionally, we note that there no longer exist single-application clusters in Paris, as distinct behaviours collapse into more general long-term patterns. We conclude that for other cities, with the exception of Periscope in two cities, spiking patterns on the day after the fire are low and few, so as to be considered noise and too difficult to parse from general random spiking to conclude an association to the fire. Several applications do not exhibit long-term behaviour past the first night, as highlighted by Cluster 5 in both Paris and in other cities.

5.3.3 Cluster Selection and Feature Values

In this section we present features values that describe the commencement time, duration, and intensity of the application traffic spikes on the day of and the day after the fire. This includes the start and end time of the each anomalous traffic spike, as well as the magnitude of deviation from baseline traffic. We also present the within cluster sum of squared error curves for the selection of k in the k means algorithm for the two-day analysis of application traffic associated with the fire.

5.3.3.1 Features, day of fire

A full description of the non-normalised values for the features of spiking applications on the day of the fire of Notre Dame is given in Tables 5.6 and 5.7. Table 5.6 shows the non-normalised values for features of applications in the city of Paris and Table 5.7 shows the values for the features of spiking applications in other cities. The distribution of the values of each feature across all applications is other cities on the day of the fire is shown in Figure 5.4.

App	s_1^{start}	s_1^{end}	s_1^{max}	$s_1^{duration}[hrs.mins]$	$s_1^{aggregate}$
Apple Video	19.15	21.15	2.10	2.00	8.42
Apple iCloud	18.45	23.45	0.40	5.00	3.54
DailyMotion	19.30	23.45	1.02	4.15	8.05
Facebook	19.00	23.45	0.32	4.45	3.52
Facebook Live	19.30	23.45	0.11	4.15	0.51
Facebook Messenger	19.15	22.45	3.54	3.30	15.29
Instagram	18.45	23.45	0.24	5.00	2.71
Molotov	19.15	21.45	2.28	2.30	9.43
Periscope	19.15	23.45	9.60	4.30	41.34
Twitter	19.00	23.45	2.93	4.45	25.95
WhatsApp	19.00	21.45	1.00	2.45	4.68

Table 5.6: Non-normalised features values for applications spiking in Paris on the day of fire.

City	App	s_1^{start}	s_1^{end}	s_1^{max}	$s_1^{duration}[hrs.mins]$	$s_1^{aggregate}$
Marseille	Facebook Messenger	19.30	22.30	0.98	3.00	3.28
	Periscope	19.15	20.45	1.95	1.30	4.83
	Twitter	19.15	23.45	0.99	4.30	7.76
Strasbourg	Apple Video	19.45	20.30	0.36	0.45	0.53
	DailyMotion	19.45	21.00	0.47	1.15	0.93
	Facebook	20.00	21.00	0.09	1.00	0.17
	Facebook Messenger	19.15	22.00	2.04	2.45	6.38
	Molotov	19.45	20.30	0.63	0.45	1.26
	Periscope	19.15	20.15	1.97	1.00	4.74
	Twitter	19.15	23.45	0.92	4.30	4.35
Lyon	Apple Video	19.45	20.45	1.04	1.00	2.49
	Facebook Messenger	19.30	21.45	1.77	2.15	5.33
	Instagram	22.00	23.45	0.03	1.45	-0.06
	Molotov	19.30	21.00	1.01	1.30	2.75
	Periscope	19.15	21.00	3.67	1.45	9.80
	Twitter	19.15	23.45	1.52	4.30	10.89
Montpellier	Apple Video	19.45	21.00	0.70	1.15	2.27
	DailyMotion	19.45	23.45	1.18	4.00	6.01
	Facebook	19.45	23.45	0.13	4.00	0.95
	Facebook Messenger	19.30	22.30	1.81	3.00	5.14
	Instagram	20.15	23.45	0.23	3.30	1.58
	Molotov	20.00	21.00	0.40	1.00	0.66
	Periscope	19.15	20.30	2.91	1.15	8.22
	Twitter	19.15	23.45	1.32	4.30	9.29
	WhatsApp	19.45	20.30	0.26	0.45	0.48
Rennes	Facebook Messenger	19.30	22.30	1.88	3.00	5.17
	Periscope	19.15	20.15	1.23	1.00	2.12
	Twitter	19.15	21.15	0.61	2.00	2.16

Table 5.7: Non-normalised feature values for applications spiking in other cities on the day of fire.

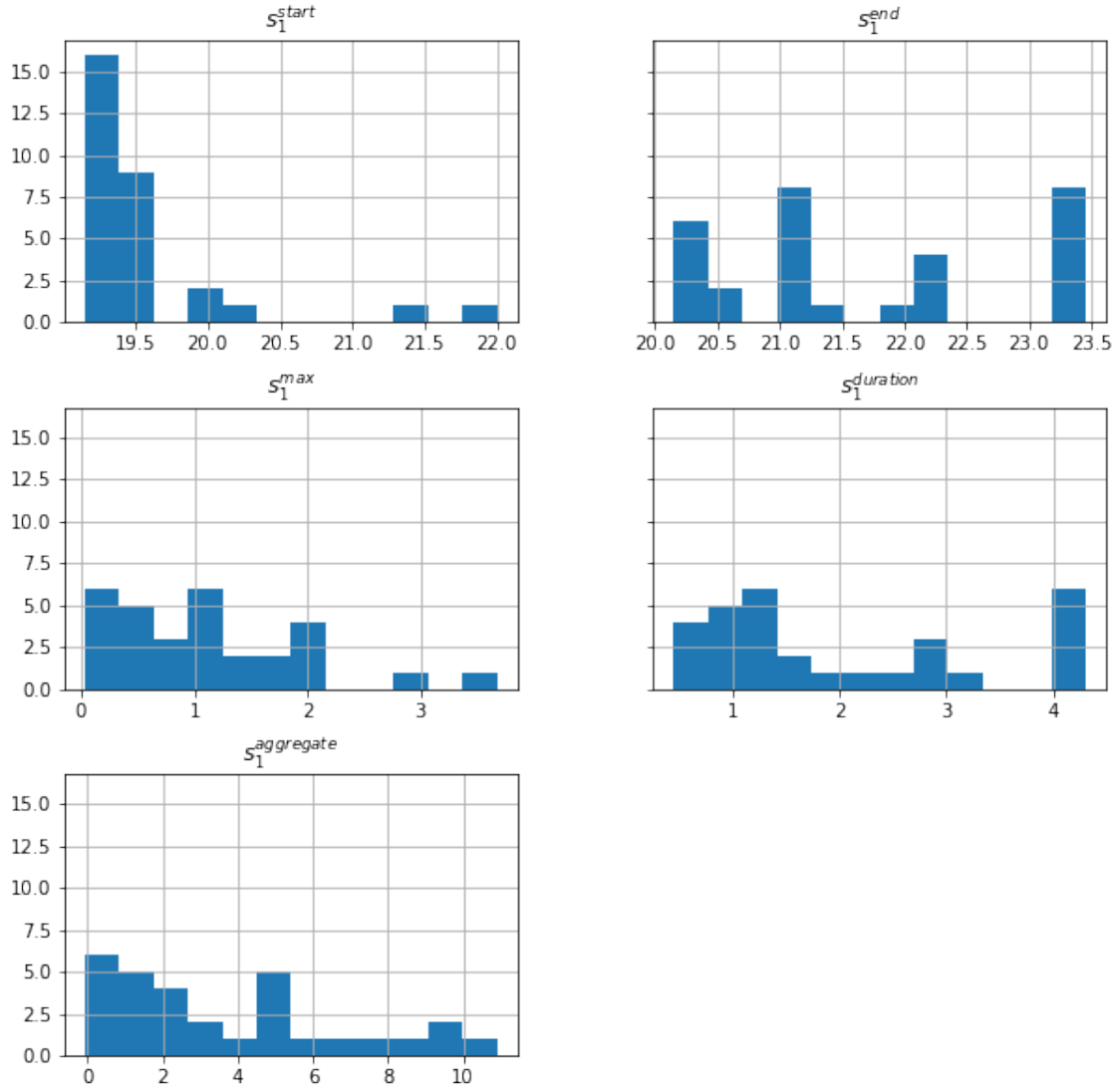

Figure 5.4: **Distributions of feature values, day of:** The distribution of the values of each feature for all other cities on the day of the fire is shown.

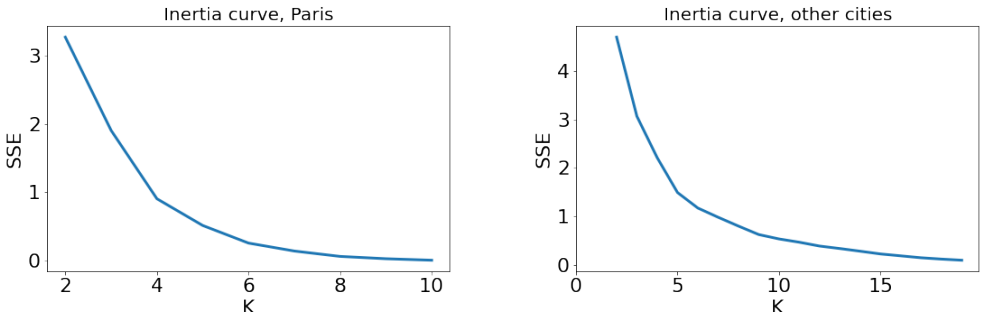

Figure 5.5: **K-means clustering, day of:** Inertia of the K-means clustering of the app-spiking features computed on the day of the fire as a function of the number of clusters K for Paris (left) and other cities (right).

5.3.3.2 Features, day after fire

A full description of the non-normalised values for the features for the day of the after fire of Notre Dame is given in Tables 5.8 and 5.9. Table 5.8 shows the non-normalised values for features of applications in the city of Paris and Table 5.9 shows the values for the features of spiking applications in other cities. The distribution of the values of each feature across all applications is other cities on the day after fire is shown in Figure 5.6.

App	s_2^{start}	s_2^{end}	s_2^{max}	$s_2^{duration}[hrs.mins]$	$s_2^{aggregate}$
Apple Video	7.00	20.45	0.57	13.45	10.60
Apple iCloud	7.15	8.00	0.07	0.45	0.10
DailyMotion	15.45	17.15	1.01	1.30	4.24
Facebook	9.00	11.45	0.03	2.45	0.18
Facebook Live	0.00	0.00	0.00	0.00	0.00
Facebook Messenger	0.00	0.00	0.00	0.00	0.00
Instagram	8.15	9.45	0.04	1.30	0.11
Molotov	7.00	16.15	1.42	9.15	14.67
Periscope	7.00	9.00	1.72	2.00	8.89
Twitter	7.00	23.45	0.33	16.45	7.88
WhatsApp	17.30	20.45	0.15	3.15	1.04

Table 5.8: Non-normalised features values for applications spiking in Paris on the day *after* the fire.

City	App	s_2^{start}	s_2^{end}	s_2^{max}	$s_2^{duration}[hrs.mins]$	$s_2^{aggregate}$
Marseille	Facebook Messenger	9.30	10.15	0.24	0.45	0.45
	Periscope	22.15	23.00	0.45	0.45	1.17
	Twitter	0.00	0.00	0.00	0.00	0.00
Strasbourg	Apple Video	0.00	0.15	0.12	0.15	0.12
	DailyMotion	16.00	17.00	0.44	1.00	1.56
	Facebook	0.00	0.00	0.00	0.00	0.00
	Facebook Messenger	0.00	0.00	0.00	0.00	0.00
	Molotov	0.00	0.30	0.25	0.30	0.47
	Periscope	21.45	23.00	1.71	1.15	3.61
	Twitter	0.00	0.00	0.00	0.00	0.00
Lyon	Apple Video	7.30	8.30	0.16	1.00	0.30
	Facebook Messenger	0.00	0.00	0.00	0.00	0.00
	Instagram	0.00	0.00	0.00	0.00	0.00
	Molotov	8.00	12.15	0.92	4.15	3.99
	Periscope	7.15	8.45	0.36	1.30	0.57
	Twitter	0.00	0.00	0.00	0.00	0.00
Montpellier	Apple Video	7.45	8.30	0.26	0.45	0.35
	DailyMotion	16.15	17.00	0.50	0.45	0.63
	Facebook	21.45	23.45	0.10	2.00	0.19
	Facebook Messenger	0.00	0.00	0.00	0.00	0.00
	Instagram	0.00	0.00	0.00	0.00	0.00
	Molotov	7.00	8.45	1.27	1.45	4.05
	Periscope	14.15	15.30	0.40	1.15	0.54
	Twitter	8.00	13.30	0.24	5.30	1.17
	WhatsApp	0.00	0.00	0.00	0.00	0.00
Rennes	Facebook Messenger	0.00	0.00	0.00	0.00	0.00
	Periscope	21.30	23.00	3.34	1.30	7.15
	Twitter	0.00	0.00	0.00	0.00	0.00

Table 5.9: Non-normalised feature values for applications spiking in other cities on the day *after* the fire.

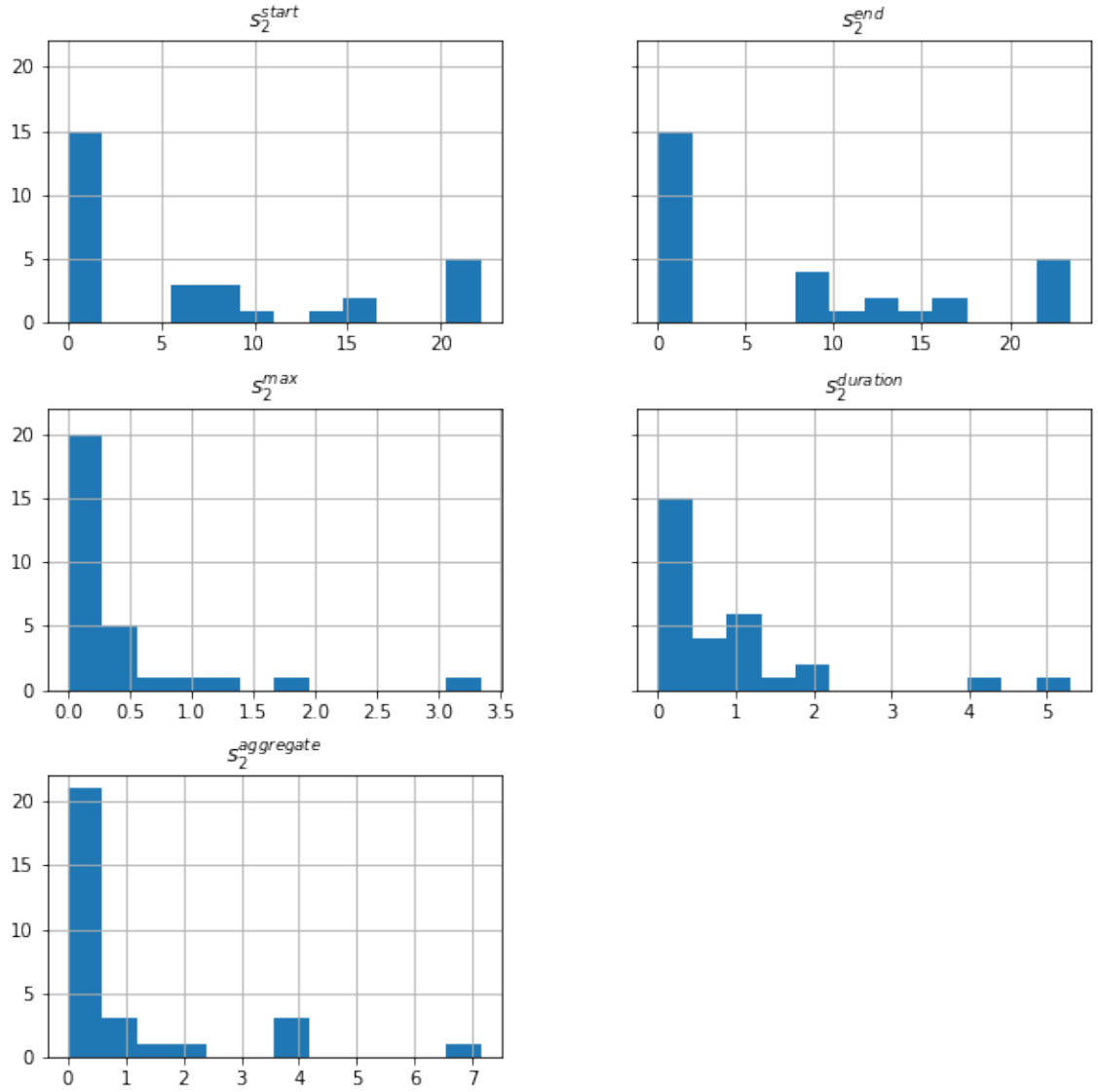

Figure 5.6: **Distributions of feature values, day after:** The distribution of the values of each feature for all other cities on the day *after* the fire is shown.

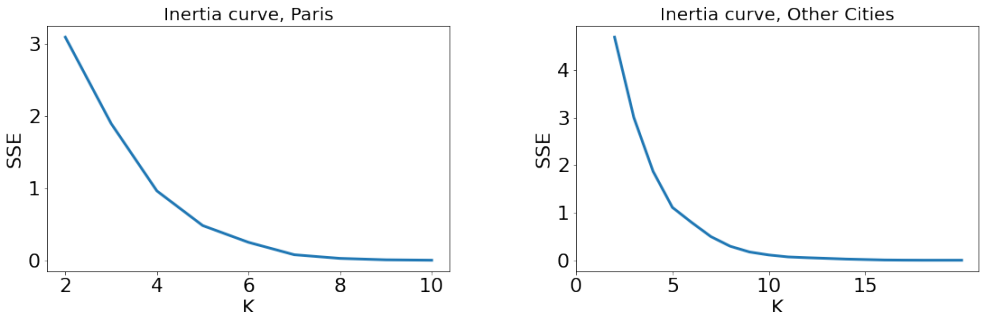

Figure 5.7: **K-means clustering, day after:** Inertia of the K-means clustering of the app-spiking features computed on the day after the fire as a function of the number of clusters K for Paris (left) and other cities (right).

5.3.4 Joint Two-Day clusters

We present the joint two-day clustering of application behaviour in response to the fire of Notre-Dame. In contrast to the analysis presented previously, which classifies short and long term behaviour of applications as separate behavioural patterns, this section considers application traffic response over an extended period of time. We define two spikes in abnormal traffic volume for consistent with our previous analysis, recovering a 10 dimensional features vector for each spiking application. This 10 dimensional vector is composed of the union of the 5 dimensional features vectors for an application across the day of and day after the fire. We perform a k-means clustering on these features vectors after normalisation, computing clustering for applications in Paris and in other cities separately, recovering 6 clusters for Paris and 5 clusters for other cities (see Figure 5.8).

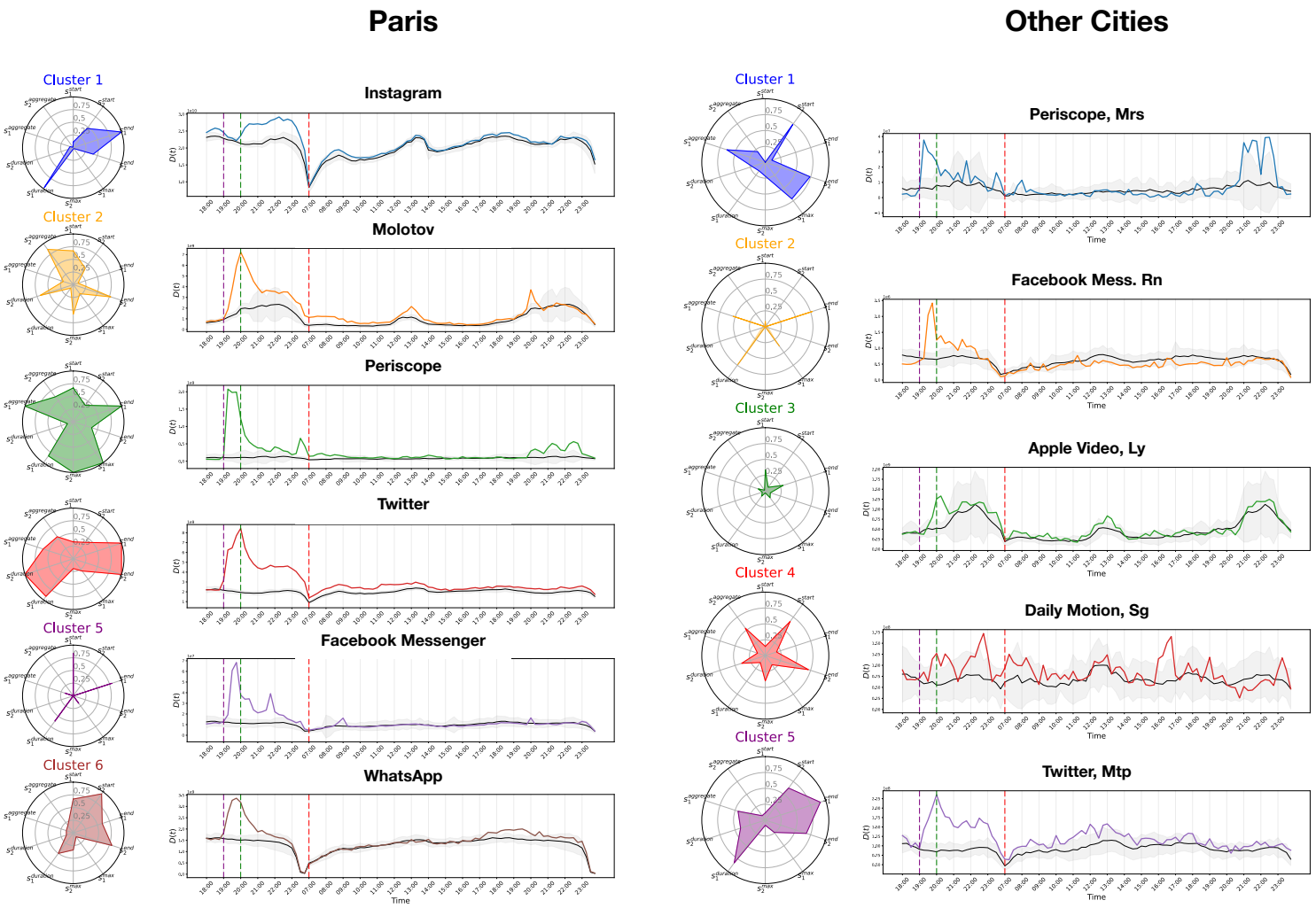

Figure 5.8: **Clusters and Representative Time Series Over Two Days: a)** For the city of Paris, the radar plots of clusters of applications are given on the left, showing the value of different features assigned to each cluster. On the right of each radar plot is a time series for an application that is representative of the applications in that cluster. Note, the information in the radar plots for each cluster is normalised within the city of Paris. The applications in each cluster are given in Table 5.10. **b)** For all other cities being considered, the radar plots of clusters of applications are given on the left, with a representative time series of an of an application in the cluster shown on the right. Note, the information in the radar plots for each cluster is normalised for all cities excluding Paris. The applications in each cluster are given in Table 5.11.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Apple iCloud Facebook Instagram	Apple Video Molotov	Periscope	Twitter	Facebook Live Facebook Messenger	DailyMotion WhatsApp

Table 5.10: Joint two-day clusters for Paris

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Periscope Mrs	Facebook Messenger Sg	Apple Video Sg	DailyMotion Sg	DailyMotion Mtp
Periscope Sg	Facebook Messenger Ly	Apple Video Ly	Molotov Ly	Facebook Mtp
Periscope Ly	Facebook Messenger Mtp	Apple Video Mtp	Molotov Mtp	Facebook Messenger Mrs
Periscope Mtp	Facebook Messenger Rn	Facebook Sg	Periscope Rn	Twitter Mtp
	Instagram Rn	Instagram Ly		
	Twitter Mrs	Molotov Sg		
	Twitter Sg	Twitter Rn		
	Twitter Ly	Whatsapp Mtp		

Table 5.11: Joint two-day clusters for other cities, Marseille (Mrs), Lyon (Ly), Montpellier (Mtp), Rennes (Rn) and Strasbourg (Sg)

In the joint two-day clusters, the short and long term behaviour of applications are considered a single pattern clustering applications by their shift in behaviour over time. From the joint clusters, we find similar trends to those seen in the two-day analysis. For example, for Paris, Twitter and Periscope both form an outlier cluster with a single defined behaviour, as is done on the day of the fire. Likewise, for other cities, we have a clustering consisting of only the app Periscope. Interestingly, for other cities the applications Twitter and Facebook Messenger, which have distinct behaviours when considered through the perspective of short and long term behaviours, cluster together when considered as a single trend in behaviour. This is largely due to both of their abnormal traffic activity being largely confined to the day of the fire.

5.3.5 Spatial Analysis of Application Traffic Response

The aim of this section is to investigate how information propagates across applications with respect to the geographic location of the source of the catastrophic event. We consider the association between information spread, in both time and intensity, and the distance from the epicentre of the catastrophe. We also aim to identify and quantify spatial patterns of above baseline traffic volume and their evolution over

time. We define spatial above-baseline traffic, $D_i(t)$, tile-wise as the difference in application traffic volume in that tile from two standard deviations above its weekday mean.

5.3.5.1 Twitter Traffic in Paris

For this analysis, we only take into account the application Twitter in the city of Paris, as it undergoes a sustained, abnormally elevated app traffic volume in the city of incidence. In Figure 5.9 we plot abnormally elevated application traffic volume, $D_i(t)$, in the metropolitan area of Paris on the day of the fire from $18:00$ to several time points following the fire. We note how, before the fire, the above-baseline traffic is spatially distributed as random noise. However, after the fire starts, we detect an approximately *radial*, outward spreading pattern emanating outward from the Notre-Dame Cathedral. Such pattern appears to persist until end of the day, relaxing over time in both intensity and *travelled* distance.

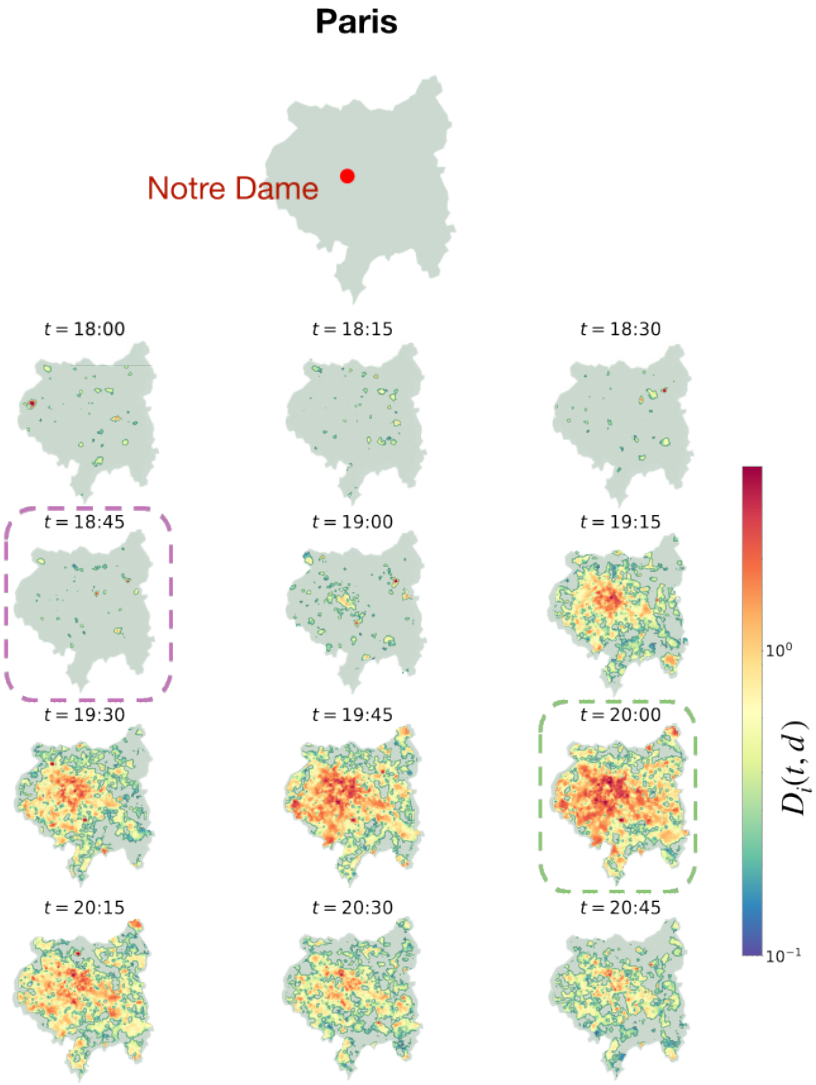

Figure 5.9: **Map of Paris with Abnormal Twitter Traffic:** The outline of the metropolitan area of Paris is shown at the top, with a red dot indicating the position of the Notre-Dame Cathedral. Below, each 15-minute timestamp shows the same city outline, now with each spatial tile as defined by the dataset displaying its corresponding tile-wise distance, $D_i(t)$, from baseline traffic. Only $D_i(t) > 0$ is considered. Recall that the first time-point considered to be related to the fire of Notre-Dame is $18:45$, outlined in purple. The time-point related to the collapse of the spire is $20:00$, outlined in green.

Given that using mobile phones, users at any physical distance could in theory access information at the same time, a pattern of radial spread that appears dependent on physical distance is unexpected. We thus quantitatively address the pattern of outward radial spread, with the cathedral as its origin. Analogously to what was done in [14], we track how the intensity of anomalous traffic, aggregated over the

course of 2 hours after the start of the catastrophe, changes with the distance from the epicentre, *i.e.*, the cathedral. We use the change in traffic volume in subsequent squares of increasing diagonal length normalised by the number of tiles within the considered distance, $\Delta D(r)$, to describe this relationship. That is, we analyse how aggregated traffic, $D(r)$, within some defined area changes as the radial distance from the epicentre increases. Noting that a radial spread of intensity of abnormal traffic outward from the epicentre would be approximated by a decaying function of this form (see Eq. 5.6), we find that in general the spread of information after the fire seems to follow a dependence on the radial distance from the epicentre (see Figure 5.10 panel a).

We can then inspect how the radial spread, or change in application traffic as a function of the radial distance from the epicentre of the catastrophe, evolves over time (see Figure 5.10 panel b). We note that in the time points before the fire, the spread of abnormal traffic spikes does not experience a dependence on the distance outward from the epicentre of the catastrophe. In the time-points immediately following the fire, there is a rapid shift in the dependency of the volume of traffic based on radius from epicentre of catastrophe (see Figure 5.10 panel b). We note that the more intense the decaying behaviour of the functions in the form of Eq. 5.6, the more the distribution of data approximates a radial behaviour. Thus, to further quantify the radial pattern of the data and how this radial pattern relaxes, we use linear regression to calculate a slope for each snapshot function shown in Figure 5.10 panel b. We can then plot how the slope of these best-fits lines change in time, with more negative slopes corresponding to more intense radial spread, which relaxes over time (see Figure 5.10 panel c).

We can also measure the KL divergence between the distribution of data in each snapshot of the form of Eq. 5.6 and the mean of the these snapshot distributions in the two hours following the fire (see Figure 5.10 panel d). This describes how each snapshot behaves with respect to the average over the two hour interval, highlighting the striking difference between the behaviour of the distributions before and following the fire, and emphasising the persistence of the data pattern in response to the fire throughout the night.

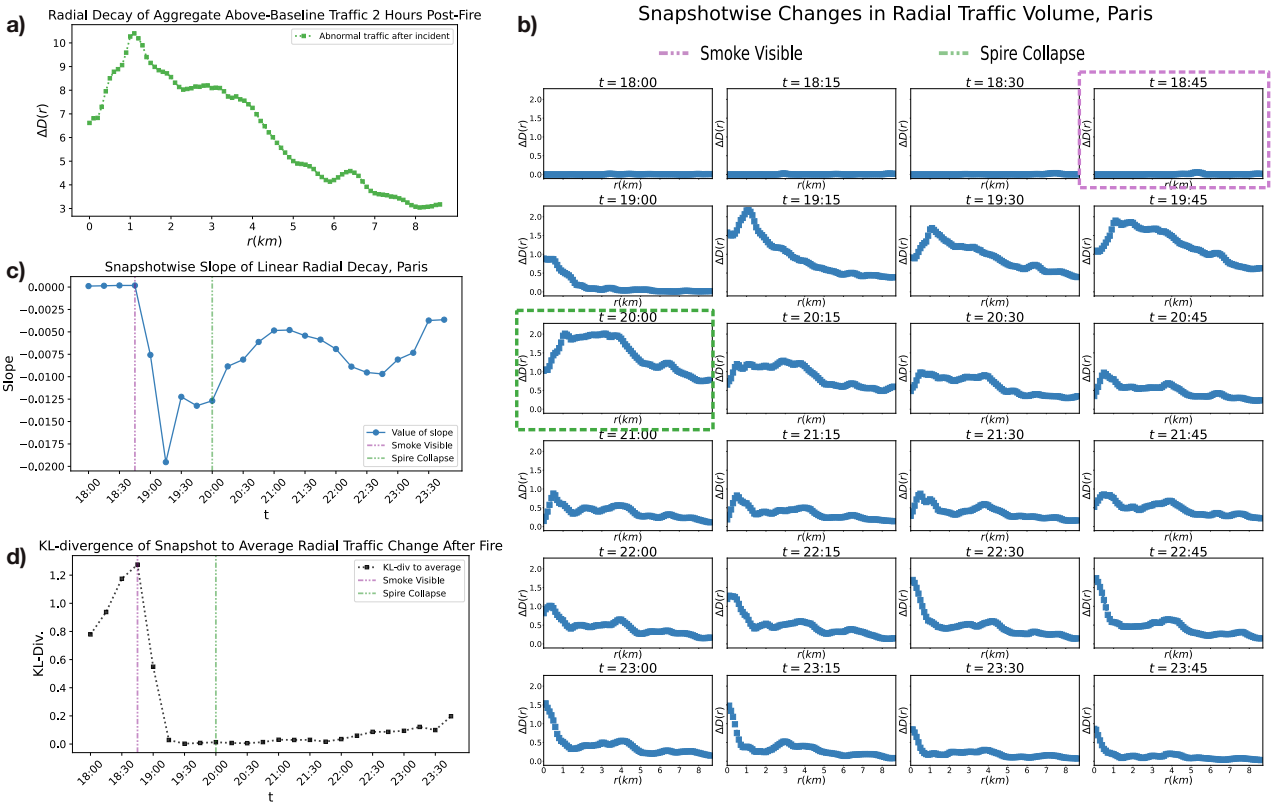

Figure 5.10: **Instantaneous and Aggregate Radial Patterns in Paris in Response to the Fire in Paris: a)** The change in abnormal traffic volume as a function of radius from the epicentre of the catastrophe (see Equation 5.6) is shown over the course of two hours (between 18 : 45 and 20 : 45), showing the course grained, general behaviour of spatial Twitter traffic response in the aftermath of the fire. The generally decreasing behaviour of the function (particularly from 1km) implies a radial spread of information. **b)** Changes in abnormal traffic volume with respect to distance from epicentre are shown at each 15-minutes time interval. From these snapshots, we can then track changes in the "radial-ity" of information spread. **c)** The slope of the best-fits line to each snapshot over time. This shows how the strength of radial spread changes over time. That is, a more radial distribution of information outward from the epicentre of the fire corresponds to a more negative slope, or a faster decay of the change in aggregate traffic from the epicentre. **d)** The Kullback-Leibler divergence of each snapshot in panel b to the average of snapshots from 18 : 45 to 20 : 45, showing how the strength of the radial spread of information varies with respect to its average change over the defined interval at each time-point. This emphasises the change in traffic spread before and after the fire and the maintaining of the radial pattern throughout the rest of the night.

We then proceed to construct a null model according to the data of abnormal traffic distribution with optimal radial spread at each snapshot in time. We can then measure the KL divergence at each time-point between the function in the form of Eq.

5.6 constructed from the data and the same functional form given by the optimally radial null model. This tells us how radial the reaction pattern in response to the catastrophe is with respect to its most radially possible distribution. We see that in the time points immediately preceding the fire, the distribution of the data is far from the distribution of the radial null model, showing a lack of dependence of abnormal traffic data from the epicentre of the fire. We then find a dramatic decrease in the divergence from the radial null model in the time points immediately following the fire, followed by a gradual rise in divergence, however, the approximation to maximum radial spread remains prevalent until the end of the day (see Fig 5.11).

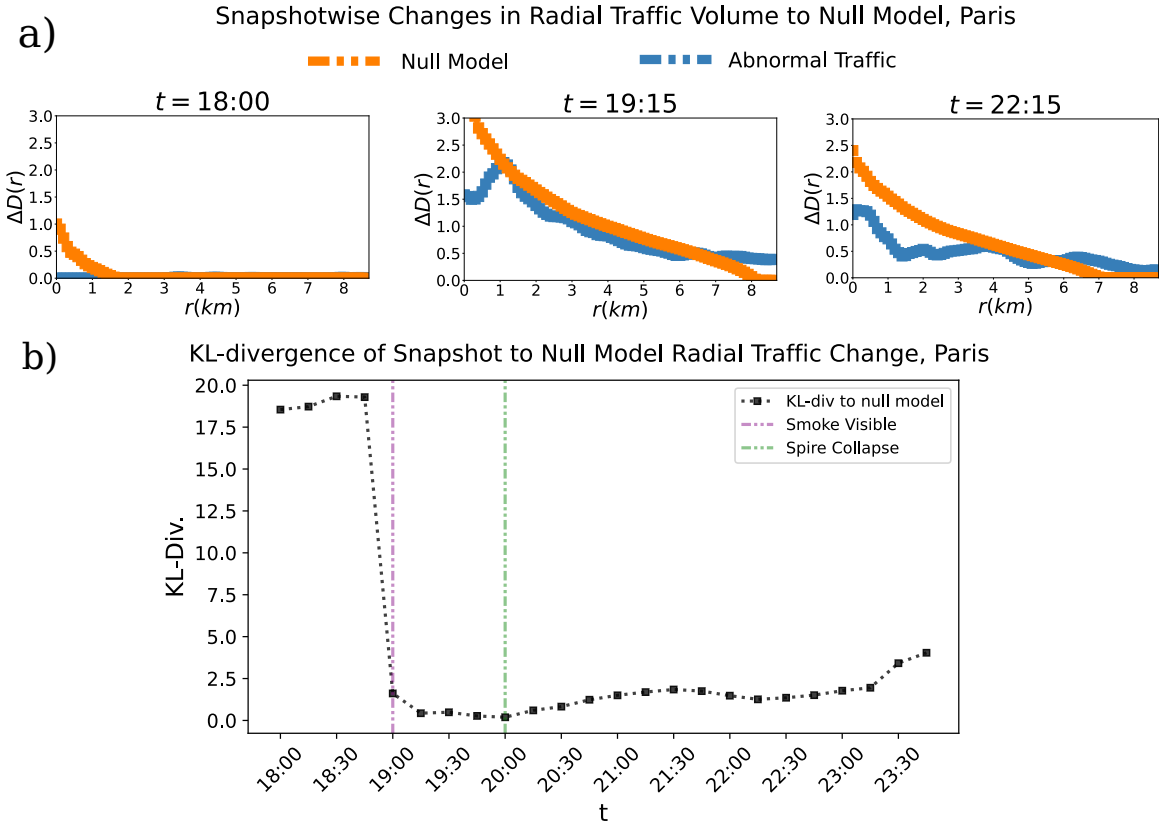

Figure 5.11: **Data vs. Null Model in Twitter Response to the Fire of Notre-Dame: a)** The change in abnormal traffic for the data and the null model at that respective time points is shown for 3 time points. Note that the first recorded time point after fire is $18:45$. **b)** The Kullback-Leibler divergence of each snapshot to the null model of most radial possible spread constructed at that time point is shown.

This is an exemplar result, as one might have previously assumed that as it is possible to receive information of catastrophe at the same time, application traffic spikes would be distributed uniformly with respect to the epicentre. However, the information and its intensity spread outwards from the epicentre, like a real-world

fire. This result is, in fact consistent with results from [14], who found a similar radial spread when looking at call traffic volume in response to catastrophic events such as earthquakes. The traffic intensity in those cases radiates outward in an radial way and the function in the form of Equation 5.6 aggregated over a two hour period hours approximates an exponential decay.

5.3.5.2 Lyon, a Further Study of Spatial Spreading Patterns

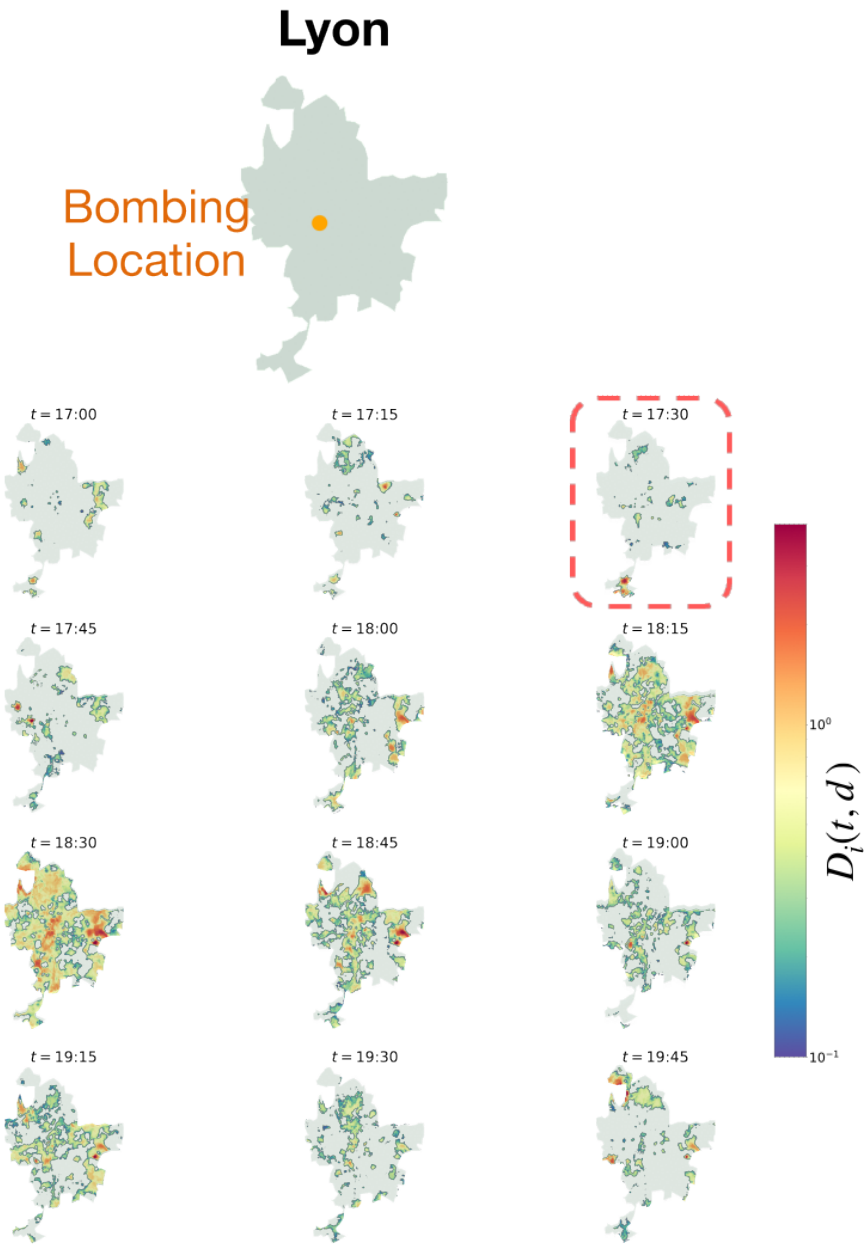

Figure 5.12: **Map of Lyon with Abnormal Twitter Traffic:** The outline of the metropolitan area of Lyon is shown at the top, with a red dot indicating the site of the bombing. Below, each 15-minute timestamp shows the same city outline, now with each spatial tile as defined by the dataset displaying its corresponding tile-wise distance, $D_i(t)$, from baseline traffic. Only $D_i(t) > 0$ is considered. Recall that the first time-point considered to be related to the bombing is $17:30$, outlined in red.

This section focuses on a second catastrophic event, a bombing in Lyon, France, occurring within the time period included the NetMob2023 dataset. We extend our

spatial analysis of spiking traffic volume of the application Twitter in response to this event in order to ascertain if there is consistency within the spatial spreading pattern results and to understand what differences may arise. The Lyon bombing occurs at approximately 17:28 on Friday, May 24th of 2019 [86]. Consistent with our analysis of the fire of Notre-Dame in Paris, we focus on the abnormal traffic of the application Twitter. We plot abnormally high application traffic volume in Lyon over several time points in Figure 5.12. Abnormally high traffic volume is measured with respect to baseline statistics calculated from all Friday's prior to the bombing within the dataset. Unlike in the Twitter traffic in Paris during the Notre-Dame fire, it is not plainly evident by inspection a clear radial spatial spread outward from the epicentre of the catastrophe. Therefore, we apply our previous spatial analysis methods to more thoroughly describe the spatial propagation of information (see Figure 5.13).

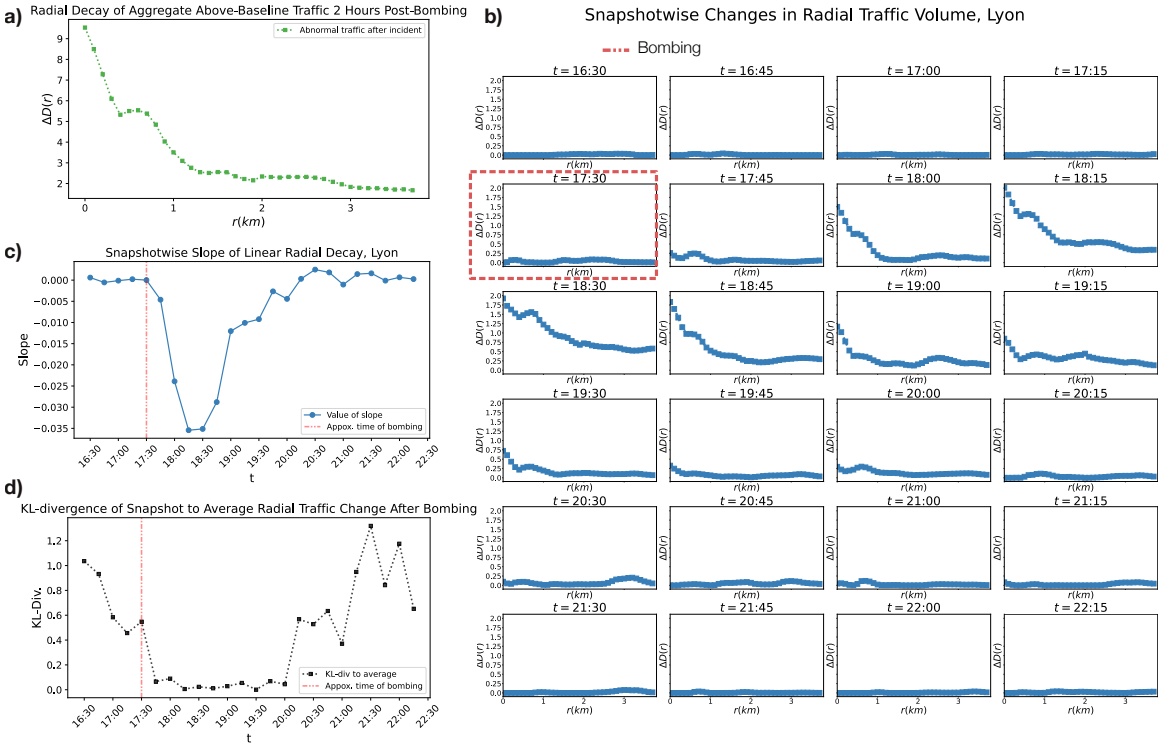

Figure 5.13: **Instantaneous and Aggregate Radial Patterns in Twitter Response to Bombing in Lyon: a)** The change in abnormal traffic volume as a function of radius from the epicentre of the catastrophe (see Equation 5.6) is shown over the course of two hours (between 17 : 30 and 19 : 30), showing the course grained behaviour of spatial Twitter traffic response in the aftermath of the bombing. The generally decreasing function implies a radial spread of information. **b)** Changes in abnormal traffic volume with respect to distance from epicentre are shown at each 15-minutes time interval. **c)** The slope of the best-fits line to each snapshot over time. This shows how the strength of radial spread changes over time. **d)** The Kullback-Leibler divergence of each snapshot in panel b to the average of snapshots from 17 : 30 to 19 : 30, showing how the strength of the radial spread of information varies with respect to its average change at each time-point.

Consistent with results from spatial methods applied to the fire of Notre-Dame, we find that in the two hours following the catastrophic event above-baseline traffic volume also experiences a dependence on the radial distance with respect to the epicentre of the catastrophe (see Figure 5.13 panel a).

We look at this pattern at each time-point to further examine the reaction of application traffic to the event and assess the excitation and relaxation of information spread. Taking functions of the form of Eq.5.6 for each time point, we can see the appearance of a radial pattern after the event, followed by a quick relaxation of this radial pattern in time (see Figure 5.13 panel b). We can quantify this as before with

slopes of linear regressions on each snapshot, once again demonstrating the evolution of the intensity of radial spread in the time following a catastrophe and its relaxation in time (see Figure 5.13 panel c). Finally, we calculate the KL divergence between the data distribution in each snapshot according to Eq.5.6 and the mean of these distributions in the two hours after the bombing (see Figure 5.13 panel d). The KL divergence of the data to the 2 hour average is far noisier than the catastrophe of the fire of Notre-Dame. While there is consistency with the mean in the time-points immediately following the fire, the duration of this consistency is far shorter that of the catastrophe in Paris. We posit that the changes to the duration of the radial spread is due to the different duration of the two events. That is, the cathedral fire occurs visibly for several hours, including the collapse of the historic spire and with continuous updates as firefighters battled the blaze. On the other hand, the event of the Lyon bombing was singular and not as long lasting.

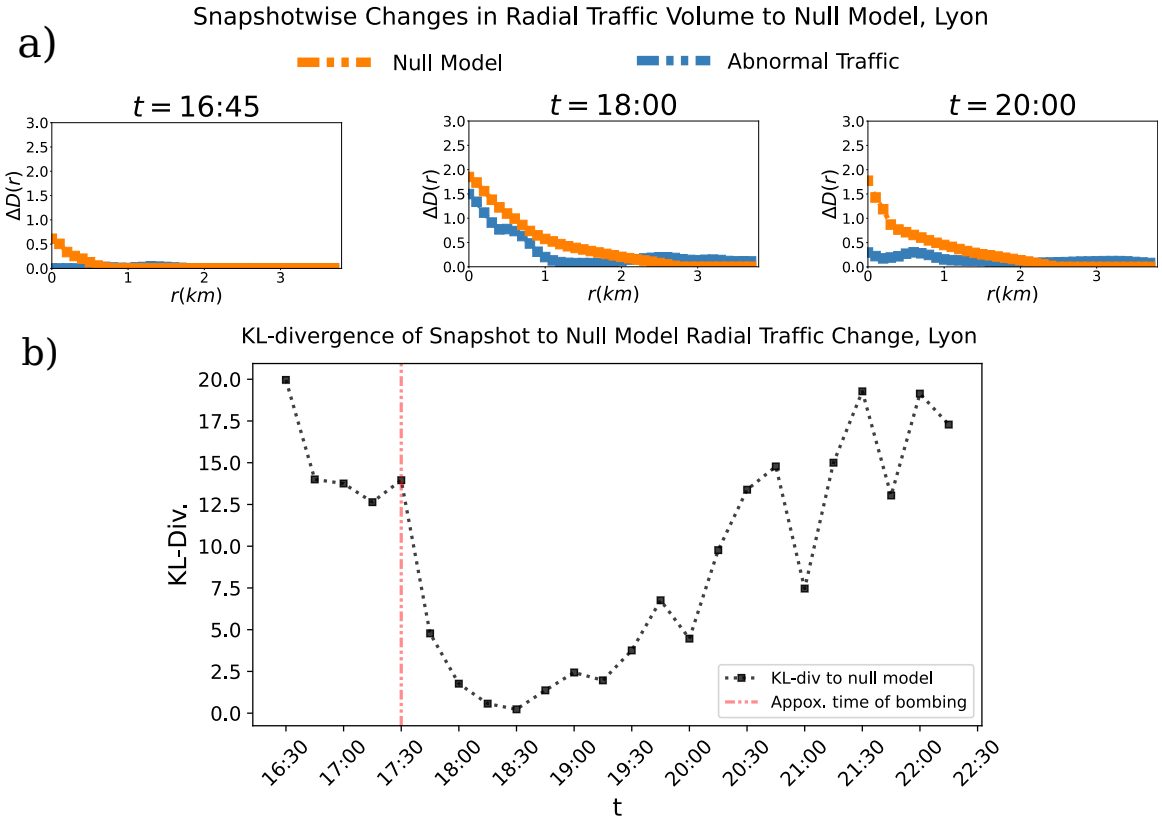

Figure 5.14: **Data vs. Null Model in Twitter Response to Bombing: a)** The change in abnormal traffic for the data and the null model at that respective time points is shown for 3 time points. Note that the first recorded time point after the bombing is 17 : 30. **b)** The Kullback-Leibler divergence of each snapshot to the null model of most radial possible spread constructed at that time point is shown.

We then construct a optimally radial null model according to previous methods and measure the KL divergence at each time point between the data and the null model to assess how closely the response pattern to the catastrophe represents a spatially radial spread (see Figure 5.14). We find that following the bombing, the data pattern closely approximates the null model of radial distribution. These results are consistent in terms of radial spread of information after a catastrophe, between the two cities. However, the different duration over which the two events unfold drastically affect the time of relaxation to the generally noisy baseline of the traffic volume after the two events.

Finally, we note that on the day of the catastrophe, there was a large planned event in Lyon, an Ed Sheeran concert, which was advertised to begin at $18:00$ [28]. In Figure 5.15, we demonstrate the extension of the methods of abnormal application traffic and pattern detection to planned events, and the interaction between two events. We note that we slightly modify our radial shell method to consider only the tiles *between* the two events *within* the city, until reaching the other event. Our results suggest a slight interaction between the traffic of the two events, which can be noted by the renewed increase in $\Delta D(r)$, following a plateauing decrease. This increase happen in the far ranges of the distance from each event, as the radius considered approaches the other event and its elevated traffic influence. However, the radial pattern is still distinguishable for several kilometres from both events given the large distance between the two events.

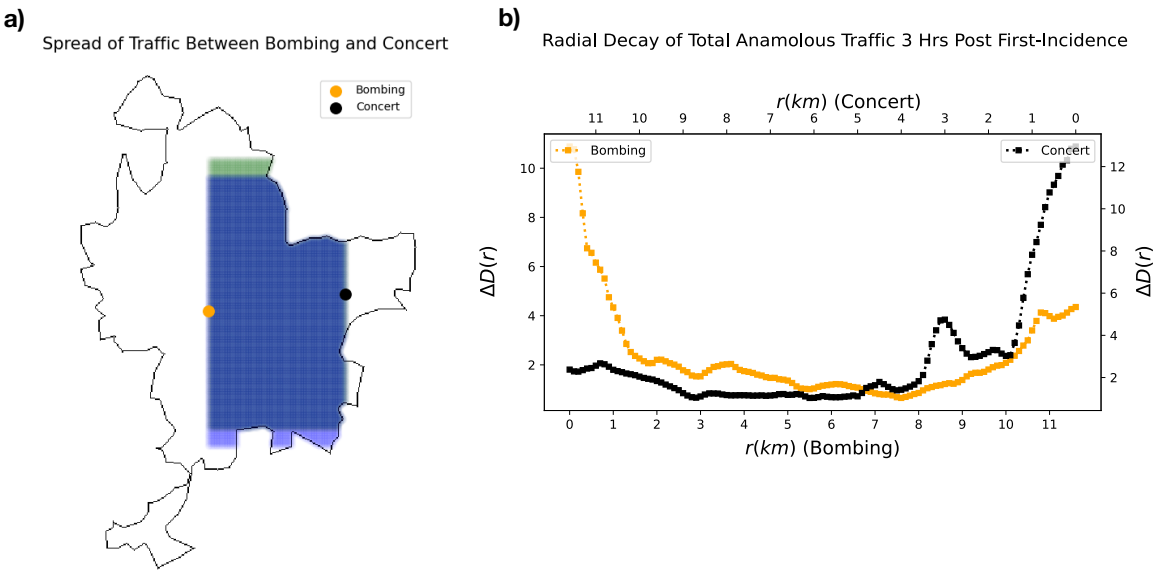

Figure 5.15: **Traffic overlap between bombing and concert in Lyon: a)** The outline of the metropolitan area of Lyon is shown, with a orange and black dots indicating the locations of the bombing and concert, respectively. The abnormal traffic between these two location *within* the city is considered, using squares of increasing radial size. **b)** The change in aggregate of abnormal traffic between the two epicentres of both the bombing and the concert are shown, from both epicentres towards the other epicentre.

5.4 Discussion

In this work we analyse time series of spatially distributed mobile phone applications and investigate the deviation from baseline activity in such traffic triggered by a catastrophic event, i.e., the fire of the Notre-Dame cathedral that took place on Monday, 15th April 2019. We use the Netmob2023 Data Challenge Dataset [144] to investigate patterns of use across applications, look into how these patterns might vary within the city of incidence as opposed to other cities, and examine how the information propagates in space and time. We are particularly interested in how temporal and spatial patterns of anomalous application traffic arise and relax over time.

We introduce a definition of *spike* of anomalous application traffic and detect the presence of such spikes for apps in the cities of Paris, Marseille, Lyon, Montpellier, Rennes and Strasbourg, on the day of and the day after the fire of Notre-Dame. We note that applications that spike in other cities are a subset of applications that spike in Paris.

The categories of apps whose time series of distance-from-baseline traffic have at least one spike include: social media, video streaming, and messaging platform. Interestingly, applications with the same function, such as social media, may exhibit differences in behaviour. For example, Instagram follows a similar behaviour pattern across all cities, characterised by short (low duration) traffic spikes across all cities. However, Twitter, another social media network, is characterised by long and sustained spikes in traffic in time (see Figure 5.2). This could be due to the tendency of users to rely on Twitter more than Instagram as a source of news and information on ongoing events [125].

To further investigate these differences in the above-baseline traffic of apps of similar function, we investigate how applications cluster together based on the spiking patterns in time over the interval of interest. We separate the day of incidence and the day after to best capture the long term behaviour of applications and how these may differ. We also perform the clustering separately for apps that spike in Paris and apps that spike in the other 5 cities to further investigate how patterns may differ in a city of incidence.

We find that between Paris and other cities on the day of interest, the applications experience generally similar trends in clustering, however, there is some diversity in the spiking patterns. For example, in application clustering can be heavily influenced by user-preference within a city. Further since broad application categories, such as social-media, cannot fully capture the diversity of application traffic response, it seems that we may need a more fine-grained look at function of applications, such as broadcasting and discussion based labelling. Distinct behaviours can be seen in Live Video streaming platforms, where short but intense spikes are seen. Also notable is the discussion based platform Twitter, which experiences a high and sustained spike. Most applications experience a fast relaxation from spiking, given that the next day many applications do not experience spikes in application traffic above baseline activity. This is especially true in cities that are not the city of incidence, showing that general interest in the event is more strongly maintained in the city of incidence.

The clustering of the time series of distance-from-baseline of app traffic in the other cities highlights how the effect of the event on an app differs across cities. Nonetheless, we note how live-streaming and, in general, video streaming apps such as Periscope, DailyMotion and Molotov, while with a delay with respect to Periscope in Paris, display a similar early spiking trend. We speculate that the delay in time of the anomalous traffic activity could be correlated with the physical distance from

Paris, as well as to the time it takes for the video and pictures of the flaming cathedral to be noticed on these platforms in other cities.

The longest lasting, most sustained spikes tend to occur in Twitter in all of the cities. Such a result suggests that Twitter might be a suitable candidate to investigate the spatio-temporal spreading of the event-triggered anomalous traffic in the city of Paris. We thus investigate the existence of a relationship between how information propagates in the virtual information sphere, and how such propagation unfolds in physical space. We find that the anomalous, above-baseline activity emanates outward of in the immediate proximity of the cathedral of Notre-Dame. We further compare the data to a null model of optimal radial spread, demonstrating the shifts to radial data patterns in response to the fire and the relaxation from radial patterns over time. This result is surprising in that the intensity of the spread appears to decay radially with the distance from the epicentre of the catastrophe, which is not immediately evident given the means of communication being investigated, that, in principle, do not rely on physical distance. This result is consistent with other studies showing that call traffic volume radiates outward from the epicentre of catastrophe [14]. We further explore the spatial spread and its relaxation through a further analysis of a bombing in the city of Lyon. We find consistent results, however, the relaxation of abnormal traffic patterns in Lyon is much quicker than the fire of Notre-Dame in Paris, likely due to the much shorter duration of the event.

There are a number of possible reasons for the radial spread effect we observe here. For example, Twitter posts are often geo-tagged, and may suggest posts to users based on their location. Furthermore, the structure of the social network itself may have a geo-spatial relationship, as social networks have been shown to be influenced by both geographic proximity and residential segregation [238]. Increased data granularity and information about the social network structure would be necessary for a deeper understanding of this radial spread effect. Finally, we show planned events may also exhibit radial spreading patterns and suggest that events may interact with each other.

In conclusion, we provide a description of how catastrophic events, particularly the fire of the Notre-Dame cathedral perturb the temporal and spatial patterns of app usage traffic. An interesting development of the current research would be to characterise the spreading of the above-baseline activity at a coarser spatial scale, thus considering entire cities as spiking tiles. This would allow us to observe spatial traffic response on the cross-city scale in relation to the event location. We posit that results regarding spatial patterns would require a more granular time resolution

than the 15 minute intervals in the current dataset. For example, Twitter traffic first spikes in Paris at 19:00, while the spikes of all other cities begin at 19:15, despite the range of a few hundred kilometres between the closest and farthest cities with respect to Paris. The other two applications that spike in all cities experience similar indistinguishable behaviours, with Periscope spiking at 19:15 in Paris and all other cities and Facebook Messenger spiking at 19:15 in Paris and 19:30 in other cities. The exception of Facebook Messenger behaviour is Strasbourg spiking at 19:15 which is in the mid-distance ranking among cities. We note that for the three applications that spike in all cities, while the maximum spike intensity for all cities is in Paris, there does not appear to be a further correlation between maximum spike intensity and city distance to Paris. To further investigate cross-city spreading patterns, we therefore would require a more fine-grained temporal scale to be able to separate the onset of the traffic spikes. Such approaches would enable the investigation of whether spatial traffic response patterns differ on intra-city and cross-city scales. Additionally, we could further investigate whether city specific user preferences regarding which applications spike also influence initial spiking times.

5.5 Conclusions and Future Directions

The analysis in Chapter 5 demonstrates how mobile phone application data can be leveraged to detect, track, and characterise patterns of information spread in response to catastrophic events both in spatially and temporally. Our findings underscore the influence of platform design on user behaviour as well as give insight into the role of city-specific user behaviour on engagement. These results contribute to the growing body of literature on digital behavioural responses to disruptive events and are of broad interest to both researchers and companies.

The development of spatial tools in Chapter 5 give insight into the flow of information to the public in response to catastrophic events as motivated by studies and interests detailed in Chapter 4. This analysis also highlights pattern of behavioural engagement in digital environments in the face of catastrophe, offering insight into how individuals interact with digital technology under conditions of physical distress.

An interesting avenue of further research would be the application of the methodology developed in Chapter 5 to the spread of misinformation. Such work could examine whether similar spatio-temporal engagement to real catastrophe patterns emerge. The application of these tools to misinformation could also illuminate how different platform features, types of content, and user-based preferences contribute to

the propagation of false or misleading information. Such studies would be impactful to platform design, content moderation, and the detection and containment of false narratives.

List of Abbreviations

KL divergence *or* **KL-div.**: Kullback-Leibler Divergence
Pa: Paris
Mrs: Marseille
Ly: Lyon
Mtp: Montpellier
Rn: Rennes
Sg: Strasbourg

Part III

Directed Hypergraph Exploration and the Chemical Space

Chapter 6

Overview and Preliminaries

6.1 Overview

Network science provides a broad framework and a set of tools for the analysis of complex systems [222, 209, 120]. Network models are powerful because of their applicability across a range of systems involving interactions between distinct components. While networks are often used to describe pairwise interactions between nodes, as exemplified in Chapter 3, they can also model sets of objects and the relationships between them, known as higher-order interactions [119, 20]. One method of modelling higher-order interactions in complex systems is to adopt a hypergraph representation, where *hyper*edges encode the interaction relations between a set of nodes. Much like their pairwise counterparts, we can also define directionality and random walk processes on hypergraphs, expanding the analytical power of network methods. Hypergraphs can be used to study phenomena across diverse systems, such as contagion processes between groups of interacting individuals [53, 213] or sets of protein interactions [154]. Hypergraphs with directionality can represent the influence of some nodes within a set on other nodes, as in the case of biological signalling networks, where directionality shows the flow of signal [108]. We say this influence flows from the *tail* of a directed hyperedge to the *head* of a directed hyperedge. As with pairwise graphs, hypergraphs allow for investigations into network dynamics and emergent properties of systems and can represent and reveal the structural organisation of systems. The use of hypergraphs and the development of hypergraph methods is an active field of investigation that is constantly evolving to understand the unique challenges of real-world systems [6].

One such application of hypergraph methodologies is towards developing an understanding of the chemical space. The chemical space is defined as the space of all chemical reactions, molecules, and their interactions and properties. The properties

of the chemical space are of interest to a variety of researchers, including chemists and network scientists [131, 187]. Explorations of the chemical space have been varied, for example, some studies have focused on the history of chemistry and the discovery of novel reactions [131], other investigations consider the topological evolution of the connectivity of molecules [71], and further efforts seek to better understand the physical world through characterisation, such as the optimal organisation of the periodic table of elements through revisions given modern chemical knowledge [130, 124]. Chemical reactions can be naturally represented by directed hypergraphs with sets of inputs of reactions mapped to hyperedge tails pointing to sets of outputs of reactions mapped to hyperegde heads, with these networks having been represented by hypergraphs in a variety of works both theoretical and applied [187, 124, 101, 74, 216]. In the classical pairwise directed network perspective, each input-output species interaction across reactions would be mapped, thereby losing the information about which specific combination of species is necessary as the multi-input of a reaction.

Chemical reactions and representations through directed hypergraphs are of interest to chemists in retrosynthesis, which is the process of identifying molecules and processes in order to synthesise a target molecule. To this end, studies of retrosynthesis have been conducted on databases of curated chemical reactions with machine learning approaches through representations as directed chemical hypergraphs [141, 232, 196, 45]. In these directed hypergraphs, a single chemical reaction is a directed hyperedge, and the set of input molecules (the tail) in the directed hyperedge points to the set of output molecules (the head) of the directed hyperedge. In this way, scientists attempting retrosynthesis can trace backwards in the graph from target outputs in a hyperedge to inputs in a hyperedge recursively. The application of hypergraph representations to the chemical space therefore unveils paths between chemical species, detailing their interactions and their relevant sequences of production.

Paths through hypergraphs can be modelled by random walks on hypergraphs. In undirected hypergraphs, these can be understood as transitions between nodes that share a hyperedge. The transition probabilities that can be defined to govern such walks are not unique. For example, the transition probability from the starting node to another node can be defined as equally split between all nodes that share an edge with the starting node, as weighted in favour of edges that have more nodes, or more likely to favour edges that were visited in the previous step [39]. In directed hypergraphs, the study of random walks is nascent, with current definitions considering a transition probability from tail to head and a transition probability associated with

moving to heads of different edges in which the tail participates [58, 67]. From these, one can introduce ideas of node ranking and community detection with the potential for rich application of these methods. The development of methods on hypergraphs derived from real-world data is of interest to the network science community as a tool for developing and testing methods and metrics for hypergraphs or directed hypergraphs as well as validating models intended to mimic real-world hypergraphs [101, 74, 180].

The following sections of this part of the thesis will be organised as follows. In sections 6.2, we present relevant concepts and methods that we use in the subsequent research chapter. In Chapter 7, we develop a methodology for walks on directed hypergraphs with an emphasis on the propagation of local connectivity and influence. This walk is motivated by the logic of chemical reactions and captures how reactivity flows and accumulates. We then detail metrics that we can derive from the proposed walk relating to the ordering, influence, and ranking of edges and nodes. We emphasise that while this methodology is inspired by chemical reactions, it is broadly applicable to systems that can be represented by directed hypergraphs. We then apply our methodology to a chemical reaction hypergraph constructed from an open-source chemical reaction dataset. We utilise the analysis of the chemical hypergraph to explore the structure and sequential ordering of the chemical space, particularly in relation to traversal of the chemical space and the influence of reactions and molecules. We compare the results of our analysis to a randomised hypergraph null model in order to draw further conclusions relating to the organisation and structure of the chemical space. Overall, we contribute a novel perspective of walks on directed hypergraphs as well as insights into the chemical space and the role of reactions and molecules.

6.2 Preliminaries

6.2.1 Hypergraphs and Directed Hypergraphs

Hypergraphs and directed hypergraphs are generalisations of pairwise graphs and directed graphs, respectively. An undirected hypergraph is a tuple $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{v_1, v_2, v_3, .., v_n\}$ is a set of nodes and $\mathcal{E} = \{e_1, e_2, e_3, ..., e_m\}$ is an associated set of hyperedges, with $e_i \subset \mathcal{V}$ for all $i \in 1 \dots m$. An undirected hypergraph is said to be *simple* if it contains no repeated hyperedges and no repeated nodes in the same hyperedge. Hyperedges describe the relationship or interaction between sets of

nodes of arbitrary size and are therefore useful in descriptions of interactions that are beyond pairwise [33, 6, 119, 180, 108, 21, 27, 181].

We further consider *directed hypergraphs*, denoted by $\mathcal{D} = (\mathcal{V}, \mathcal{E})$. As before, $\mathcal{V}$ is a finite set of nodes but $\mathcal{E}$ is now a finite set of *directed hyperedges*. Each hyperedge $e_i \in \mathcal{E}$ is of the form $e_i = (T, H)$, where T and H are subsets of $\mathcal{V}$ and encode that the *tail* of the edge T points to the *head* of the edge H. We consider a directed hypergraph to be *simple* if it contains no repeated edges and no repeated nodes within the same edge. This further implies no self-loops, where self-loops are defined by the existence of a node in both the head and the tail of an edge.

Directed hypergraphs are a natural mathematical representation for chemical reactions, as the inputs and outputs, or reactants and products, of a reaction define the tail and head of a directed hyperedge [108, 187, 74, 141, 196]. For instance, the chemical reaction

$$A + B + C \longrightarrow D + E,$$

where A, B, and C are the inputs of the reaction and D and E are the outputs of the reaction, defines an edge $e_1 = (t_1, h_1)$ such that $t_1 = \{A, B, C\}$ and $h_1 = \{D, E\}$.

6.2.2 Molecular and Synthetic Complexity

Another notion relevant to our study of the chemical space is that of "molecular complexity". Molecular complexity, or the quantification of how complex a molecule is in terms of structure, is widely regarded to be of significance in the synthesis of organic chemistry as well as drug discovery and planning [169, 158, 37, 235]. Molecular complexity is an often debated topic in the scientific community, with no firm consensus of what constitutes a good metric for measuring complexity [169, 158]. Many indices regarding the quantification of molecular complexity have been proposed under a variety of techniques [235]. For example, the size of molecules [37], their molecular topology [184], the information content of molecular graphs [25], the sensitivity to stereochemistry and symmetry [37], and the number of chiral centres of a molecule [231, 19] have all been considered in measures of molecular complexity.

Strictly molecular complexity approaches, however, have some drawbacks, particularly in the view of synthetic complexity. For example, there exist simple reactions that can produce complex structures, as in the case of condensation reactions. Further, easily available complex reagents allow for the introduction complex features with ease and some structurally complex molecules can be prepared in a single step [63, 78].

Synthetic complexity, in contrast to molecular complexity, aims to capture the ease with which a target molecule can be synthesised [235]. This generally relates to how many synthetic steps are required to synthesis a target molecule [235, 63]. The notion of synthetic complexity is therefore akin to notions of path length through connecting reactions. Defining the synthetic complexity of molecules has been paid vast interest, with computational approaches developed to consider chemical and molecular properties [94, 46, 100] and other approaches using step-based methodologies [46, 96]. Step-based methodologies in particular, such as the SCScore for synthetic complexity, take into account paths and chemical properties in combination with deep-learning approaches. We note that approaches like this are pairwise in terms of associations of molecules.

Later in this thesis, our focus will be motivated by path-based approaches to understanding synthetic complexity. Contextually, we do not explicitly derive formalised synthetic complexity metrics *per se*, however, we inform and complement knowledge pertaining to the development of synthetic complexity and frame synthetic complexity in the context of network connectivity.

Our hypergraph based random walk and the associated metrics we introduce investigate the propagation of reactivity, the capacity for enabling reaction and molecule production, the potential of chemical species to be produced, and the positioning of molecules and reactions in the framework of the wider chemical space through the lens of network science. Our methodology is therefore are well situated in the context of path-based approaches to synthetic complexity. The methodology we develop allows us to navigate the connectivity of the complex chemical space and therefore, shed light on it.

6.2.2.1 Assembly Theory

The general field of assembly theory is also relevant to this part of the thesis. Assembly Theory is a stepwise methodology that seeks applications in retrosynthesis and focuses on least-step paths [201, 129]. Assembly theory understands objects as constructive under a combinatorial framework and seeks shortest paths of development of these objects. Similar to the methodology we develop, assembly theory is step-wise, combinatorial, and considers objects repeatable once they have been discovered.

Despite these similarities, we note key distinctions between assembly theory and our developed methodology. Firstly, our methodology is a system-wide, exploratory method that is meant to capture the growth of possibilities in a system. Our method aims to quantify how richly a system can be expanded and in what depth, breadth,

and ordering this can be done. Our methodology is developed in the framework of a network context and contains the notion of ordering and alternative pathways. This aims to situate the organisation and the reachability of reactions and nodes across a system-wide chemical context. Assembly theory focuses its aims on the building and constructiveness of specific objects and has a particular focus on the minimal steps for construction, as opposed to the wider properties of systems. Assembly theory is focused on irreducibility as opposed to reachability potential. This also leads to the collapse of multiplicity, as opposed to our methodology and developed metrics where multiplicity is framed as informative.

Chapter 7

Cascade Processes in Directed Hypergraphs and the Architecture of the Chemical Space

7.1 Introduction

In this Chapter, we introduce a novel framework for random walks on directed hypergraphs based on the availability of nodes and guided by principles pertaining to the reactivity of chemical compounds. Directed hypergraphs allow us to capture multi-way interactions, and the development of methodologies on directed hypergraphs is an active area of investigation within complex systems. Our methodology unveils the organisation and structural properties of these systems. We further introduce graph-theoretic metrics on the nodes and edges of the hypergraph based on the application of our novel methodology, which unveil key drivers of the hypergraph system. Our metrics develop the *reactive capital* of nodes and edges in relation to their capacity for production, the ease with which they can be produced, and their general association to other nodes and edges in the directed hypergraph.

As a playground, we extend our investigation to the application of modelling of chemical reactions, capturing the properties of organisation within the chemical space. We compare the results of our methodology, and in particular the metrics derived for nodes and edges, between a chemical reaction dataset and a randomised null model of this data. This analysis illustrates how our method can be applied and also allows us to outline properties of the chemical space and its connectivity.

Hence, in this chapter, we present a broad graph-theoretic approach to complex systems analysis. The methods we develop introduce a novel approach to modelling connectivity in directed hypergraphs, as well as offer a network-based perspective of

the organisation and traversal of the chemical space.

7.2 Reactive Cascade and Reactive Walks

In this section, we develop our methodology for the analysis of connectivity in directed hypergraphs. The development of our approach is inspired by chemical reactions and captures the reactivity process. To characterise how nodes affect each other indirectly, through chains of reactions, we define a notion of connectivity on directed hypergraphs. We note that this descriptive methodology may also be applied in general to hypergraphs beyond chemically-derived hypergraphs. We term the method that we develop the *reactive cascade*.

A reactive cascade is defined as follows. We select a starting hyperedge, $e_{initial} \in \mathcal{E}$, and consider the set of all nodes $v \in e_{initial}$ to form the set of available nodes η. Given a set of available nodes, we consider another edge $e = (T, H) \in \mathcal{E}$ to be available if all nodes in the tail of e are any subset of nodes in η, i.e. $T \subset \eta$. We denote the set of available edges as $\Psi_E(\eta)$ and the corresponding set of available tails as $\Psi_T(\eta)$. We then simulate forward the reactions of all available edges, which corresponds to adding its head nodes to the set of available nodes and adding the edge to the set of reacted edges, $\mathcal{R} \subset \mathcal{E}$.

Intuitively, this process simulates a sequence of reactions occurring as input species become available. This process can be repeated until there are no edges whose tails are a subset of available nodes, including the case where all possible edges have been visited. We can conceptualise the notion of the reactive cascade as the spreading of the occurrence of all possible reactions as chemical species are accessed through reaction. We note that this method is not meant to account for the rates of reaction of species, only the possibility of reaction. Since the reactive cascade is the result of forward simulating chemical reactions starting from a single reaction, it defines the set of reactions and chemical species that are reachable from an initial reaction, or any set of input species. At each step of the reactive cascade, the next step is sampled from the power set of available nodes, $\mathcal{P}(\eta)$. We note that multiple edges may be reacted at any step, as it is possible that multiple tails may be subsets of $\mathcal{P}(\eta)$. This leads to the potential for many more combinations in the power set in successive steps, or *levels*, of the cascade. For this reason, we describe this process as a *cascade*. Further, at any level within the reactive cascade, we may only sample new edges and tails corresponding to $\mathcal{P}(\eta)$ as determined by edges in the set $R \cup e_{initial}$ in

the current level. Therefore, the order in which edges are added to R determines the evolution of the reactive cascade.

The reactive cascade can also be seen as a snowball sampling with edges being added by sampling level, s_n, as reactions occur [82]. For example, consider a single product reactions in the chemical space with a reactive cascade on the edge e_0. The reactions pertaining to this example, as well as a visualisation of the reactive cascade are shown in Figure 7.1. Suppose e_0 corresponds to the reaction

$$e_0 : A + B + C + D \longrightarrow E.$$

The available nodes are $\eta = \{A, B, C, D, E\}$. Suppose that the reactions that are now available, or equivalently the edges whose tails are in $\mathcal{P}(\eta)$, are now

$$\begin{aligned} e_1 &: A + C + D \longrightarrow G, \\ e_2 &: A + B \longrightarrow F, \\ e_3 &: D + E \longrightarrow H. \end{aligned}$$

Therefore, $\Psi_E = \{e_1, e_2, e_3\}$ and we denote the first sampling level of the cascade as $s_1 = e_1, e_2, e_3$. We then assume a reaction on all possible available tails, or $R = \{e_1, e_2, e_3\}$. Continuing our example, we update $\eta = \{A, B, C, D, E, G, F, H\}$ and we recover the next reaction level, s_2, from $\mathcal{P}(\eta)$, composed of the reactions

$$\begin{aligned} e_4 &: B + F \longrightarrow J \text{ and} \\ e_5 &: D + H \longrightarrow K, \end{aligned}$$

Such that $\Psi_E = \{e_4, e_5\} = s_2$. We take another step after updating $R = \{e_1, e_2, e_3, e_4, e_5\}$ and $\eta = \{A, B, C, D, E, G, F, H, J, K\}$, with $\Psi_E = \{e_6\}$ from

$$e_6 : F + J \longrightarrow L.$$

Finally, after updating η and R, we can suppose that in this example there are no more edges such that the edge tails are a subset of $\eta = \{A, B, C, D, E, G, F, H, J, K, L\}$ and $\Psi_E = \emptyset$. Thus, the reactive cascade for e_0, $C(e_0)$, is composed of three levels given by

$$s_0 : e_0$$

$$s_1 : e_1, e_2, e_3$$

$$s_2 : e_4, e_5$$

$$s_3 : e_6.$$

The reacted edges in $C(e_0)$ are $\{e_0, e_1, e_2, e_3, e_4, e_5, e_6\}$ and the visited nodes are $\{A, B, C, D, E, F, G, H, J, K, L\}$. Algorithm 1 considers a reactive cascade on an initial edge.

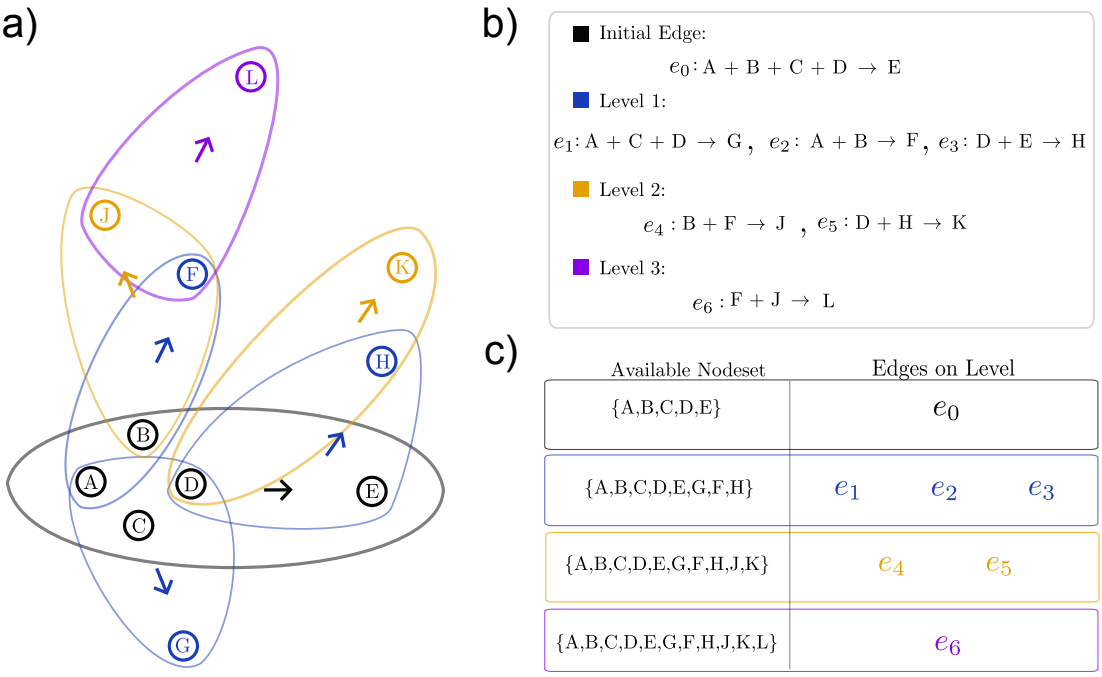

Figure 7.1: **Reactive Cascade Example:** A visual representation of a reactive cascade beginning with initial edge e_0. All panels are colour coded to correspond with each other. **a)** A visual representation of the single product reaction of chemical reactions that are reached in the reactive cascade is shown. Arrows point from the tail of the hyperedge to the head of the hyperedge. **b)** The sampling levels of the reaction cascade, with edges and their corresponding chemical reactions are displayed. **c)** The available nodeset, η, and the corresponding edges added on each level are shown.

In a reactive cascade, we consider reaction upon all available edges. We can define a *reactive walk* as a sequence of edges sampled from a reactive cascade. That is, instead of considering the reaction of all available edges at a time, we select a single edge whose tail is a subset of $\mathcal{P}(\eta)$ and then react and update based on the addition

of this edge. We emphasise that a reactive walk from a hyperedge is necessarily a sequence of edges sampled from the reactive cascade of that edge, not all samples are viable reactive walks. For the above example, a reactive walk from edge e_0, $W(e_0)$, could be the sequences $W(e_0) = e_0, e_2, e_4, e_6$ or $W(e_0) = e_0, e_3, e_5$. Note that the latter sequence does not contain e_6. $W(e_0)$ could not be the sequence $W(e_0) = e_0, e_1, e_5$, as node H is not available, and $W(e_0)$ could not be the sequence $W(e_0) = e_0, e_3, e_4$, as node F is not available. However, walks need not be constructed by sampling level or sampling levels could be revisited, for example the sequence $W(e_0) = e_0, e_2, e_4, e_1, e_3, e_5, e_6$ could be considered valid.

Algorithm 1 Conduct an exhaustive reactive cascade search starting on some initial edge.

procedure REACTIVE CASCADE($C(e)$)
    Choose $e_0 \in \mathcal{E} \subseteq D$
Initialise tracked quantities
    $\mathcal{R} \leftarrow e_0$ ▷ Set of reacted edges
    $\Psi_E(\eta) \leftarrow \varnothing$ ▷ Set of available edges
    $s_0 \leftarrow \{e_0\}$ ▷ Cascade level 0
    $i \leftarrow 0$ ▷ Level counter
    $\eta \leftarrow v,\ v \in t \subset e_0$ ▷ Set of available nodes
    Define $\eta_{new} = \eta$
    while $\eta_{new} \neq \emptyset$ **do**
        for each node $v \in \eta$, and each tail T containing v **do**
            if $T \subseteq \eta$ and related edge not reacted, $e = (T, H) \notin \mathcal{R}$ **then**
                Update $\Psi_E(\eta)$ with all edges whose tails are subsets of η
            end if
        end for
        Update $\eta_{new} = \bigcup_{e \in s_n} H(e) - \eta$
        Define cascade level $s_i \leftarrow \Psi_E(\eta)$
        React to all available edges: $\mathcal{R} \leftarrow \mathcal{R} \cup s_i$
        Update available nodes:

$$\eta \leftarrow \eta \cup \bigcup_{e \in s_n} H(e)$$

        Update level: $i \leftarrow i + 1$
    end whilereturn reactive cascade
end procedure

We note that the reactive cascade algorithm need not be initialised on a defined on an edge $e \in \mathcal{E}$, but can be applied to any subset of nodes in $\mathcal{V}$. This would then describe the reactive cascade process as spreading from this set of nodes as the initial

available nodes. For example, we can describe a reactive cascade on the molecules that are involved in the most number of reactions and analyse the resulting reactive cascade. In these cases, supposing we denote our chosen node set as $v_{selected}$, we initialise $\mathcal{R} = \emptyset$ and $\eta = v_{selected} = s_0$ and proceed with Algorithm 1.

While we have developed the reactive cascade and reactive walk in the context of chemical reaction hypergraphs, we note that the methodology detailed here is also applicable to other areas and hypergraph applications. For example, in gene regulatory networks sets of genes regulate other genes and the activation of upstream genes is required for the activation downstream genes, lending the system to the application of the spreading methodology developed as the reactive cascade. Further, we can see similarly organised systems in signal transduction, where the specific ordering of signals transmitted to cells is required, as well as in task scheduling, parallel computing, and puzzle solving, where certain tasks are required to proceed with others.

7.2.1 Properties of the Reactive Cascade

So far, we have detailed a methodology investigating the relationship between nodes and hyperedges, with a focus on ordering in walks and graph traversal. Further motivated by chemical objects, we proceed to detailing properties of the edges and nodes with the purpose of capturing the "reactive capital" of nodes and edges, or molecules and reactions. Towards this end, we investigate the position of these objects in the reactive cascade and their potential for unlocking further reactions. We will first characterise the properties of the edges, or reactions, followed by analogous properties of the nodes, or molecules. These properties form an investigative set that can reveal meaningful characteristics of objects or relations within the system. However, they do not represent an exhaustive catalogue of all possible properties that might be derived from the reactive cascade across different applications. We distinguish between edge-based and node-based metrics, presented below.

- **Reaction Metrics: Edges**

  Total Elements in Cascade

  The most basic way to characterise the role of an individual reaction in the larger chemical reaction hyperedge space is by the total elements in the reaction cascade that is generated by using the reaction as the starting edge of the cascade. This gives a description of how many other reactions can be reached in the reaction space, or how much of the chemical space can be pervaded, from

this edge. Considering the example from Figure 7.1, the total edge elements in the reactive cascade from e_0 is 6, from the six edges reached in the reactive cascade.

Number of Levels in Cascade

A second way to characterise the role of a reaction is by the total number of levels in the cascade. This gives a description of how many times the reaction step, equivalent to the completion of the valid reactions from available nodes, is undertaken. That is, the number of levels cascade contains information on the amount of reactive steps that are taken in order to move through the chemical space. In contrast to the total elements in the cascade, which describes the breadth of reactions that can be reached, the number of levels in the cascade captures the sequence in which reactions can be added. For example, an edge can have many elements in its cascade but few levels, implying that the essential molecules for the permeation of the chemical space are resampled from existing available molecules rather than from the addition of new edges. Alternatively, an edge that has few elements and many levels in its reactive cascade demonstrates that in order to traverse the chemical space, more intermediate steps are required and the addition of edges is more ordered. In Figure 7.1, e_0 is characterised by 3 levels.

Edge Profile

The profile of an edge (reaction) is a description of how edges are sampled in its corresponding levels. The edge profile combines the information given by the total number of elements reached in a reactive cascade and the number of levels in that cascade. An edge profile, EP, for any edge is given by a vector of the length of the maximum level for any cascade across the entire dataset. Each entry in the vector is the number of new edges reached in the reactive cascade at that level, starting from the edge being profiled. If there are no new edges added in a level of reactive cascade, the entries in the vector at the position of that level are zero. Suppose for some dataset, the initial edge e_{max} is the edge whose reaction cascade reached the most levels and initial edge selection e_{random} does not. Example edge profiles could be EP(e_{max}) = [3, 5, 7, 1, 2] and EP(e_{random}) = [2, 1, 0, 0, 0]. In this case, e_{max} has 18 total edges added to its 5-level cascade, distributed into the corresponding vector, and e_{random} has 3

total edges added to its 2-level cascade, with two edges at level one and one edge at level one. Taking the cascade from Figure 7.1, the edge profile of e_0 is [3, 2, 1].

Total Number of Cascade Appearances

The total number of appearances of an edge indicates how many times across all cascades that reaction is produced in the chemical space from any other initial starting edge. This pertains specifically to the how many times in the dataset the edge can be reacted into and does not include the use of the edge as the initial starting point. Edges with higher cascade appearances have more production routes or possibilities, while edges with lower cascade appearances have fewer options that lead to the production of that reaction. Taking the cascade from Figure 7.1, the total number of cascade appearances of edges e_1 to e_6 increases by one due to the reactive cascade from e_0.

Edge Reactive Out-degree

This metric defines a notion of out-degree on edges within the reactive cascade. We note that within the reactive cascade, the reaction step of the cascade which adds the head of the edge to the set of available nodes is what drives the further percolation of the cascade. The reactive out-degree of an edge therefore is defined as the number of times the head nodes associated to an edge are sampled from the set of available nodes in the tails of edges in previous levels of the reactive cascade. That is, the reactive out-degree of an edge tracks how many new reactions in the cascade an edge is essential to producing. As an example, suppose there are three edges in level 1 of a reactive cascade, e_1, e_2, and e_3. Suppose the head node of e_2 is an input node for one reaction in level 2 and the head node of e_3 is an input node for one reaction in level 3 and two reactions in level 4. Supposing also the head node for e_1 is not sampled as an input node in subsequent levels. We would then say that the reactive out-degree of the edges e_1, e_2, and e_3 are 0, 1, and 3, respectively. Note that if a head node of an edge in the cascade already appears in the initial edge, it is not considered a new node that the reaction of an edge has unlocked. Therefore, input nodes that sample this head node in a subsequent level will not add to the reactive out-degree of the cascade edge. The reactive cascade shown in Figure 7.1 increases the edge reactive out-degree of edges e_2, e_3, and e_4 by one.

Average Previous and Subsequent Levels or Elements

The average number of total previous and subsequent levels of an edge considers the position of an edge in relation to the production steps of the cascades it participates in. Similarly, the average number of previous and subsequent elements of an edge considers the total amount of edges in the previous and subsequent levels of the cascade the edge participates in, focusing more on the size of the cascades. The previous and subsequent levels or elements deduces positional information of the edge in relation to other edges. Taking edge e_5 from the reactive cascade shown in Figure 7.1 as an example, the number of previous and subsequent levels is both 1. The number of previous elements is 3 and the number of subsequent elements is 1.

Mean Appearance Level

The mean appearance level of an edge is the average level in which the edge reacts across all reactive cascades in which the edge appears. Conceptually, this captures how early or late in cascade steps the reaction occurs, regardless of how often the reaction occurs. For example, an edge that consistently appears in the first levels may be easily produced, which may be the case whether or not the reaction is commonly produced. Likewise, an edge that appears towards the end may be more difficult to produce in that it requires more reactions to achieve. In combination with other metrics such as reactive out-degree or the total number of subsequent elements, this metric is indicative of how essential an edge is in unlocking new parts of the dihypergraph, or accessing swathes of new reactions. This is particularly relevant for datasets with large cascades. Taking the example from Figure 7.1, the appearance level for e_5 from the cascade from e_0 is 2.

Average Elements of Cascade Participation

An edge can be characterised according to the average number of elements of cascades the edge participates in, i.e. cascades in which the edge is reacted into. Edges that occur in larger cascades are associated with more structurally complex reaction subspaces, while edges in smaller cascades are associated to more isolated subsections of the chemical reaction hypergraph. Taking the example from Figure 7.1, edges e_1 to e_6 have 6 elements in this cascade they are participating in.

- **Reaction Metrics: Nodes**

  Cascade Elements as Initial Node

  A node can characterised by the total number of other nodes in all reactive cascades from which the node participates in the initial selected edge. In essence, this metric describes the set of all nodes which can be accessed through reactions that are associated with this node. Considering Figure 7.1, the initial nodes A,B,C,D and E all add six nodes to their cascade elements from nodes F,H,G,J,K,and L.

  Number of Levels in Cascade as Initial Node

  A node may also be characterised by the average number of levels in reactive cascades produced from reactions which the node participates in the initial selected edge. This shows how many reactive steps initial reactions associated to this node take in order to complete their cascade. In the example in Figure 7.1 the initial nodes A,B,C,D and E are associated with three levels in the cascade from e_0.

  Total Node Sampling Profile

  The total node sampling profile of a node (molecule) is a description of the use of that node within reactions reached in the reactive cascade at different levels. That is, if a node is in the tail of an edge reached within a reactive cascade, we say that the node is "sampled" by that edge in its corresponding cascade level. This node may be sampled by more than one edge on the same level. We construct the total node sampling profile by assigning each node a vector of the length of the maximum level any node can be sampled in across the entire dataset. Each position in the vector is then filled with the sum of times the node is sampled at that corresponding level across all cascades in which it is sampled. Entries in the vector where the node is not sampled are zero. For example, supposing the final node sampling maximum is at level 4, a node sampled once at level 1 and twice in level 2 in a cascade and the same node sampled three times in level 1 and once in level 3 in a different cascade will have the total node sampling profile TNSP($node$) = [4,2,1,0]. This gives a distribution indicating the importance of the node in accessing new reactions across cascade steps. The total node sampling profile for node A in the example from Figure 7.1 is [2,0,0] and for node D is [1,1,0].

Weighted Sampling Node Appearance

We construct a weighted sampling node appearance score using information based on the total node sampling profile. The weighted appearance score for a node is given by the sum of its sampling appearances multiplied by the corresponding level of that appearance. This can be expressed by

$$\sum_{n=1}^{n_{max}} n(f_n),$$

where n_{max} is the maximum levels and f_n is the number of sampling appearances of the node at that level. Taking our previous example with total node sampling profile TNSP($node$) = [4,2,1,0], the weighted node appearance score for this node would be

$$(1 \times 4) + (2 \times 2) + (3 \times 1) + (4 \times 0) = 11.$$

The weighted node sampling appearance for node A in the example from Figure 7.1 is 2 and for node D is 3, indicating that node D is more connected depth-wise than node A.

Production Appearances

This metric characterises a node by the total number of times the node is produced across all cascades. This indicates how many starting reactions can give rise to the same chemical species. Nodes with higher production appearance scores have more potential production routes, while with lower scores have fewer production options. This metric differs from the sampling appearance metric in that it tracks how many times the node is produced rather than utilised. The appearance of a node is unique per cascade, that is once a node is produced from a reaction step within cascade, a different reaction in the same cascade producing this node is not considered an additional appearance. In the example in Figure 7.1, the nodes G,F,H,K,J, and L are all produced once in the cascade from the reactive cascade starting from e_0.

Head Out Degree

This metric tracks the number of times a node that is produced in the head of a reaction is used as an input in any subsequent level. In essence, this tracks how essential this head node is for subsequent reaction steps. This is summed across all reactive cascades the node participates in to retrieve the total number of new reactions that rely on the use of this produced node, analogous to

the edge out-degree metric. In Figure 7.1, nodes F,J,and H increase their head out-degree by one in the reactive cascade from e_0.

Average Previous and Subsequent Levels or Elements

Analogous to the edge metric, the average number of total previous and subsequent levels or node elements of a node reveals positional information regarding a node in relation to other nodes in the cascades in which it participates. This is measured by node first appearance in the reactive cascade. Taking as an example node F in Figure 7.1, the number of previous levels is 0 and the number of subsequent levels is 2. The number of previous elements is 0 and the number of subsequent elements is 2 (nodes J and L).

Mean Appearance Level

The mean appearance level of a node is the average level in which the node first appears across all reactive cascades in which the node appears. This may indicate how difficult a node is to produce in terms of reactive steps required within reactive cascades. Similar to the edge metric, we may consider this metric in combination with head out-degree and total number of subsequent elements to indicate whether a node is essential in unlocking new parts of the graph. Taking as an example node H in Figure 7.1 the reactive cascade from edge e_0 contributes an appearance at level 1.

Average Elements of Cascade Participation

The average number of node elements of cascades the node participates indicates how intertwined this node is with other areas of the chemical hypergraph. In the example in Figure 7.1, nodes F,H,G,K,J, and L all add six elements to this metric from the reactive cascade from edge e_0.

7.3 Application

In the previous section, we have presented the reactive cascade methodology and the associated reaction metrics to capture the reactive capital of edges and nodes through their involvement in reactive cascades. We now move to an application of the reactive walk and metrics to a chemical reaction hypergraph derived from data. As discussed in Chapter 6, there is an interest in categorising and ranking the importance of reactions and molecules, as well as in describing pathways through chemical datasets.

This application is exploratory in nature and gives an alternative description of reactive capital from a network-based approach. We begin by describing the chemical dataset and our data processing steps in order to build a chemical hypergraph based on the available dataset. We then describe general properties of the chemical hypergraph based on the reactive cascade and the previously described metrics. We compare results on the original chemical hypergraph and on a randomised hypergraph in order to identify structural properties of the chemical space.

7.3.1 Dataset and Data Processing

We collected chemical reaction data from the open source, organic chemical reaction database repository, the Open Reaction Database (ORD) [105]. The Open Reaction Database contains datasets curated by scientists of single-step organic reactions. Over a million reactions have been uploaded to the database.

We processed extracted datasets from the ORD using ORDerly [232] and canonicalised the molecules in the dataset reactions using RDKit [122] in Python. In the construction of the chemical hypergraph from the datasets, each node is a chemical species that is identified by a canonical SMILES (Simplified Molecular Input Line Entry System) string, which encodes the three dimensional structure of the chemical into a string of characters [228]. The directed hyperedges of the chemical hypergraph are specified reactions.

The role of a chemical species within any chemical reaction is categorised by the ORD as a reactant, agent, solvent, or product. These chemical roles are defined by the user input of the ORD, presumably researchers uploading their reactions to the database. We consider the roles reactant, agent, and solvent as input roles (corresponding to tails) and the role product as the output role (corresponding to heads). We note that a single chemical species can occupy multiple roles across different reactions. Our specific focus is to give a general overview of combinations of reactants that lead to outputs, noting that agents and solvents are auxiliary given their function is to facilitate these reactions. Our cleaning of this dataset reflects this purpose, however, other choices of exhaustive cleanings of this dataset may entail highly involved machine learning projects.

We have cleaned the dataset as follows. We begin by removing from the output chemical species that are also present in the input, as these correspond to leftover species that have not been fully depleted during the chemical reaction. That is,

$$A + B + C \longrightarrow D + C,$$

becomes

$$A + B + C \longrightarrow D.$$

We then limit the reactions under consideration to single output reactions, to investigate explicitly chemical reactions that seek to optimise the production of a single target molecule during synthesis. This step also aligns with literature concerning machine learning embeddings that filter chemical datasets to single product reactions [232, 141]. Our goal in describing the chemical space is to give a simple representation of the combination of chemical species inputs that produce chemical outputs. As such, we remove species categorised as agents and solvents from the datasets as these are participants that facilitate reactions but are not part of the product of the reaction. Hyperedges therefore contain input reactants in their tails and output products in their heads. We then ensure that there are no repetitions in input species as these are considered redundant. We also remove single input and zero input reactions, as these do not explicitly describe interactions between reactant species. We further remove all reactions that contain an invalid SMILES encoding or SMILES reaction encoding, as we consider these to be user errors. Finally, we remove duplicate reactions from the directed chemical hypergraph. Our cleaning steps produce a simple directed hypergraph.

We note that without removal of agents and solvents, the resulting hypergraph contains an order of magnitude more hyperedges and nodes than the fully preprocessed hypergraph. Before preprocessing, the directed chemical hypergraph with the inclusion of agents and solvents contains 1.7 million hyperedges and 1.2 million nodes. After preprocessing, the directed chemical hypergraph contains 362,341 hyperedges and 585,069 nodes. This change has significant implications on the connectivity of the hypergraph derived from the ORD. In particular, the reactive cascade application resulting from a construction including agents and solvents has 18 levels, while the preprocessed hypergraph, which will be presented in the following sections, has 5 levels. This indicates that spreading patterns with agent and solvent inclusion have substantially more sequential ordering. We emphasise that the dataset may be cleaned under different choices and our methodology of reactive cascades may be applied just the same. However, an investigation with different choices as to which species to include in the chemical hypergraph is outside of the scope of this work. It is also indicated within the data that there is user mislabelling of chemical species, particularly by the relatively high number of input species in reactions following the data cleaning. However, more thorough cleaning of user data is a separate machine learning problem that requires an inspection of molecules as well as an advanced

knowledge of chemistry. In this work, we choose to analyse the data as is, using a few simple pre-processing steps as an illustration of our methodology. In future work, it may be fruitful to explore more detailed data cleaning and pre-processing.

7.3.2 Directed Hypergraph Null Model Randomisation

The directed hypergraph null model was generated using the methodology described by [180]. The so-called NUDHY-JOINT randomisation preserves the in- and out-degree of nodes as well as the tail and head size associated to each of those nodes. The randomisation is performed via a Markov chain Monte Carlo method.

7.3.3 Results

In the following subsections, we present the application of our directed hypergraph methodology to the chemical reaction dataset from the ORD. We apply our methodology by initiating a reactive cascade on each reaction, or edge, individually following our cleaning and preprocessing steps. That is, each edge is treated as the seed, or the initial edge, of the reactive cascade across all possible edges. The edge and node metrics described result from this seeding. We then compare the results of the reactive cascade and associated metrics between the data and the randomised null model. We also demonstrate the application of the reactive cascade on a selected subset of nodes from the ORD to demonstrate the versatility of our methodology. The details of the subset of these nodes selected is given in Section 7.3.4.

7.3.3.1 Hypergraph Statistics

From the available datasets recovered from the ORD following data processing, we construct a directed chemical hypergraph with 585,069 nodes, or molecules, and 362,341 hyperedges, or reactions. The number of unique products is 343,170, meaning there are 19,171 or 5.29% of the products produced by more than one reaction. Further, the percentage of reactants that appear as products is 14.56%, with the percentage of products that appear as reactants at 12.01%. The molecules that participate in both roles is 7.05% of all molecules. This is suggestive of cycles in the graph and provides information to the chemical community, as it demonstrates the potential for paths for traversing the graph and producing target molecules via multiple routes during retrosynthesis.

We initially investigate the connectivity of the chemical hypergraph by its undirected component structure. This is equivalent to the concept of weakly connected

components in a pairwise graph. Undirected connected components are constructed by removing the directionality of the hypergraph. A hyperedge is then considered connected to another hyperedge if they have any node in common. We observe that the chemical hypergraph is *not* strongly connected. The chemical hypergraph does exhibit a largest connected component (LCC) which contains 92.68% of the nodes and 95.58% of the edges of the undirected chemical hypergraph. The connectivity of the graph is surprising, considering there is little overlap between head and tail molecules. This suggests that chemical researchers reuse known reactants in the synthesis of products. This is consistent with the literature on the history of the exploration of the chemical space, indicating that scientists tend to select compounds that are readily available and well-known to their expertise from which to extend their inquiries [131]. In Figure 7.2 we show the appearance of reactants by order of magnitude as well as the percentage of all reactions that include the most frequent reactants. We can see that significant percentages of the reactions contain the top reactants and while there are many novel reactants introduced in the dataset, these are used in combination with known reactants.

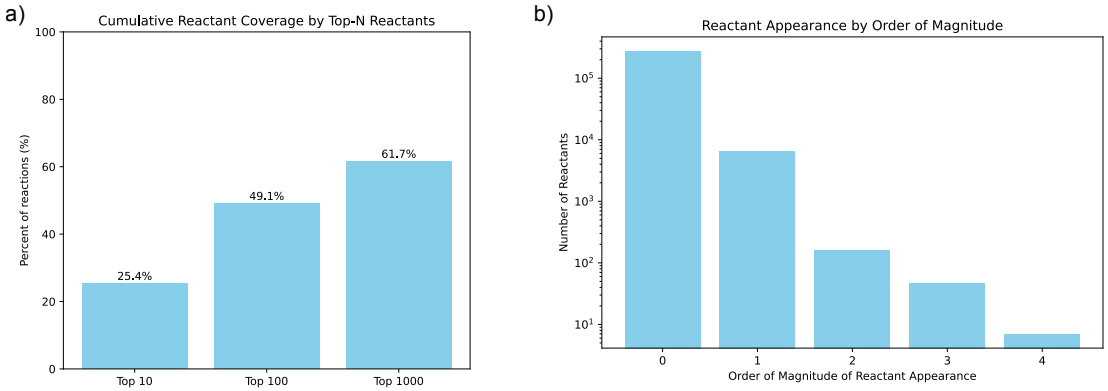

Figure 7.2: **Information on Reactants: a)** The percent of reactions that include at least one of the most frequently appearing reactants (inputs). **b)** The frequency of reactant appearance by order of magnitude. We note that there are 283,126 unique reactants.

The size distribution of all connected components, excluding the LCC, is shown in Figure 7.4 panel a. We can see that the sizes of the remaining connected components are very small in relation to the LCC. These remaining connected components may be representative of smaller, more exploratory undertakings and again emphasise the intuition that chemists generally operate within the largest undirected component of the chemical reaction space.

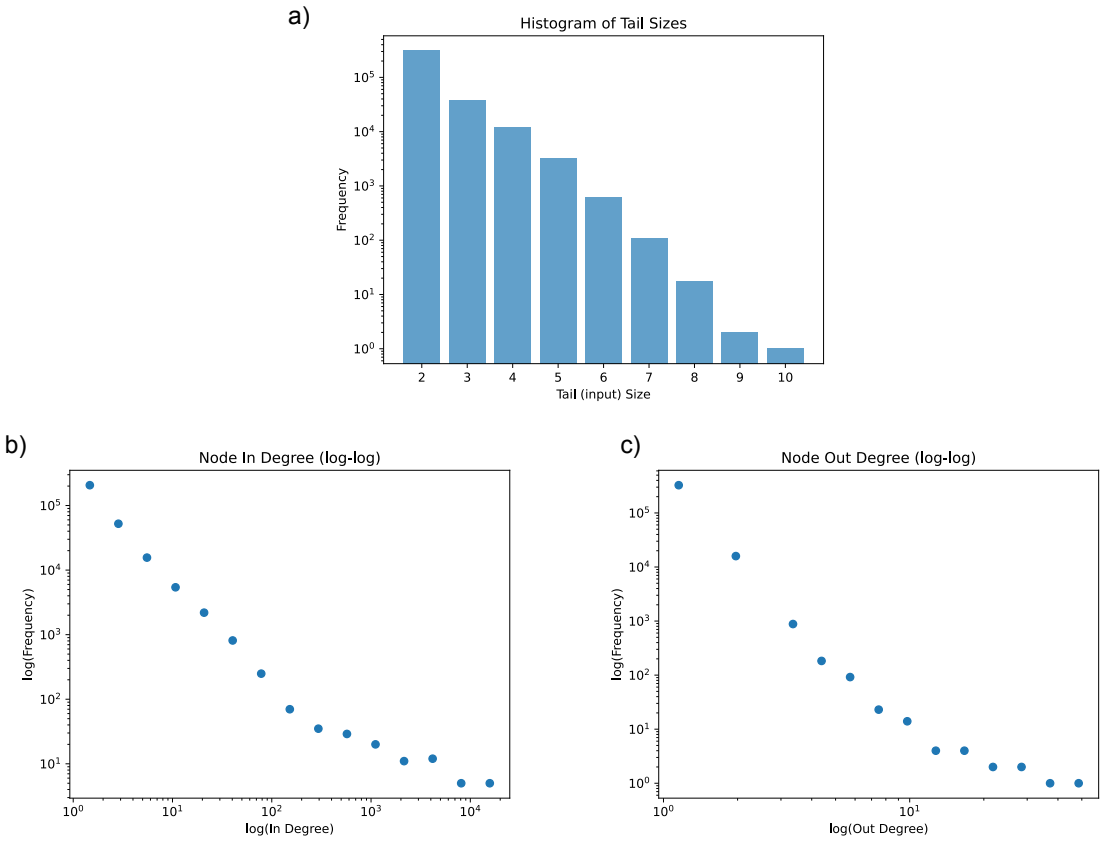

Figure 7.3: **Details of the chemical space: a)** Frequency of tail sizes, or number of reaction inputs for the chemical hypergraphs. We note that the number of inputs can become quite high for synthesis. Having many inputs to a reaction is an indicator of user input misclassification of reactants in the ORD, however the detailed investigation of this is beyond the scope of this work. **b)** The in-degree of nodes on a log-log scale. **c)** The out-degree of nodes on a log-log scale. We note that the log-log plots of node in- and out- degree denote a heavy tailed distribution.

We present further details regarding the organisation of the chemical space through the distribution of the tail sizes of the hypergraph, corresponding to the number of inputs in the reactions. We also demonstrate the in- and out- degree of the nodes, corresponding to the frequencies at which nodes are in tails or heads of a hyperedge respectively. This is presented in Figure 7.3. With regards to the in-degree, we see that most nodes are produced few times, with a small amount of nodes being produced many times. With regards to out-degree, we see that similarly, most nodes are produced by few other nodes, with a small amount of nodes produced by many nodes.

We generate a null model randomisation of the chemical hypergraph. As expected,

the null model randomisation will preserve the tail size as well as the in- and out-degree distribution of the nodes. The LCC of the randomised model contains 90.6% of the original number of nodes and 94.8% of the original number of edges. The difference in number of nodes and hyperedges in the randomisation and the original data is minimal. In subsequent sections, we will reveal that in spite of this there are considerable structural differences between the two. The distribution of size of all other connected components excluding the LCC is shown in Figure 7.4 panel b.

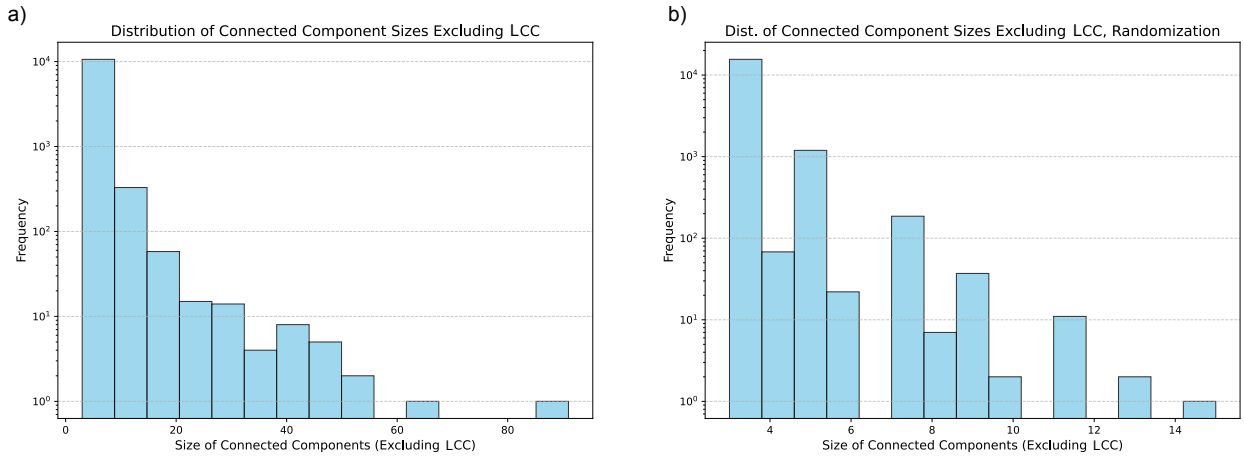

Figure 7.4: **Data v. Randomised Undirected Connected Component: a)** Size of undirected connected components for the chemical hypergraph derived from data. **b)** Size of undirected connected components for the hypergraph from the randomised null model.

As a next step, we apply the reactive cascade methodology to both the data derived hypergraph and the randomised model hypergraph. In the following sections we will detail the results of the applications of our derived metrics of the reactive cascade on the chemical space. In the application of the reactive cascade to the dataset, 16.26% of the reactions, or initial edges, react into, or walk, to least one more reaction. For all other reactions, no other edge can be reached as a subset of the nodes of that initial reaction. In the randomised null model, 14.95% of initial edges react into at least another edge.

7.3.3.2 Edge Metrics

In Figure 7.5, we demonstrate the application of our proposed metrics for reactive capital on edges in the data recovered from the ORD (left hand side panels) and the randomised null model hypergraph (right hand side panels) following the reactive cascade on both hypergraphs. This information gives insight into the structure of

reactive steps needed for reactions and the general traversal of the chemical space. Our comparison with the null model demonstrates that the metrics are not merely random, but rather can lead to meaningful interpretation of the data from the ORD. The metrics are presented in the order in which they are introduced in Section 7.2.

We begin by noting that while the number of edges of the chemical hypergraph is on the order of magnitude of 10^5, a single reactive cascade can reach at most an order of magnitude 10^3 edges in a reactive cascade. This promotes the isolated nature of reactions, suggesting that the chemical space cannot be thoroughly explored through individual edge reactive cascades, or further, that a single reaction is limited in its ability to connect to the wider chemical space. Further, the size of the space of all edges that can be reached by a reactive cascade from any initial reaction is 32,868 or 9.07% of all reactions. We note that this is indicative of a highly disconnected chemical reaction space upon the inclusion of directionality in the hypergraph. This contrasts with the fact that there is a large connected component in the undirected case. This lack of connectivity in the directed case again implies that there are limited capabilities in the traversal of the graph. We note that our starting points under consideration are selections of a single reaction with a single product produced at each iteration. For the randomised data, the size of the space of all edges that can be reached by a reactive cascade from initial any reaction is 26,428, or 7.29% of all reactions.

The most prominent difference between the reactive cascade on the ORD data and the randomised null model is the maximum number of levels in the cascade, five levels in the data and three in the randomised model. This indicates that the connectivity in the random model is more breadth based while the reactions in the ORD require more sequential ordering and iterations to progress through the chemical space. This difference is a consequence of chemically relevant information impacting the chemical data.

In Figures 7.5a) we note the number of edges the reactive cascade reaches across all initial edge selections. While the distribution of edges reacted into in the chemical reaction data indicates slightly more edges reached, the distribution of edges reacted into in the randomised model is broadly similar. Hence, the level structure is key in understanding the differences in the two spreading processes, indicating that while edges may be similarly reached, the pathways by which spreading occurs differ substantially and not otherwise obvious.

We show the distribution of levels and the number of times an edge propagates to that number of levels in Figure 7.5b). In Figures 7.5c) we show the sum of the

edge profiles of all edges of the data and randomisation. This shows that both in the reaction data and the randomisation, reachability narrows as levels proceed. This is not readily apparent *a priori* as another plausible scenario would be a widening of possible available edges as more nodes are added to the available nodeset. In Figure 7.5d) we show the frequency of the appearances in cascades of edges as a participant being reacted into. We can see that for the chemical reaction data, edges are more likely to appear in a cascade than in the randomised data. This is reflective of underlying chemical information, indicating that there are reactions are more likely to be repeated in the cascades. These may be common reactions, important transition reactions in connecting the chemical space, or reactions with frequently appearing molecules.

In Figure 7.5e) we show the edge reactive out-degree of the two models. We first note that the edge out degree distribution of the chemical reaction data contains an order of magnitude more than the randomised data. This is due to the influence of more levels in the chemical reaction data leading to more possibilities to resample a previous available edge. Notably, fewer edges in total are influential for unlocking subsequent reactions in the chemical reaction data than the random data, with 3.83% edges sampled in the addition of a new edge to a cascade in the data and 5.47% of the edges sampled in new cascade edges in the random data. This pertains to information on which chemical reactions are key in traversing the graph. Figures 7.5f) and ff) show the distribution of the average previous and subsequent levels of all edges in the models. These distributions are shown as averages in order to capture the system-wide trends. However, we emphasise that for an individual reaction, the average previous and subsequent levels of an edge reveals information about its position in the cascade. That is, the edge may be indicated to be towards the start or end of long or short cascades, or any possible intermediate position, reflecting its role in the reaction sequence. In Figure 7.5g) we show the mean appearance level of an edge. We also emphasise how many of the edges do not appear, or are not reacted into, in the cascade. The mean appearance shows the expected value of edge positioning across the system.

Finally in Figure 7.5h) we present the average elements in a cascade that an edge participates in across the systems. We emphasise that for individual reactions, this is indicative of how intertwined reactions are with other reactions in the chemical space. For example, a reaction may not react into many other edges in its own cascade but may frequently occur through interactions with other edges and different cascade starting points. We find examples of such reactions in the chemical data, as well as

examples of edges in the data that react widely into other edges when chosen as the starting point of a reaction cascade, but are not themselves produced when starting at other edges.

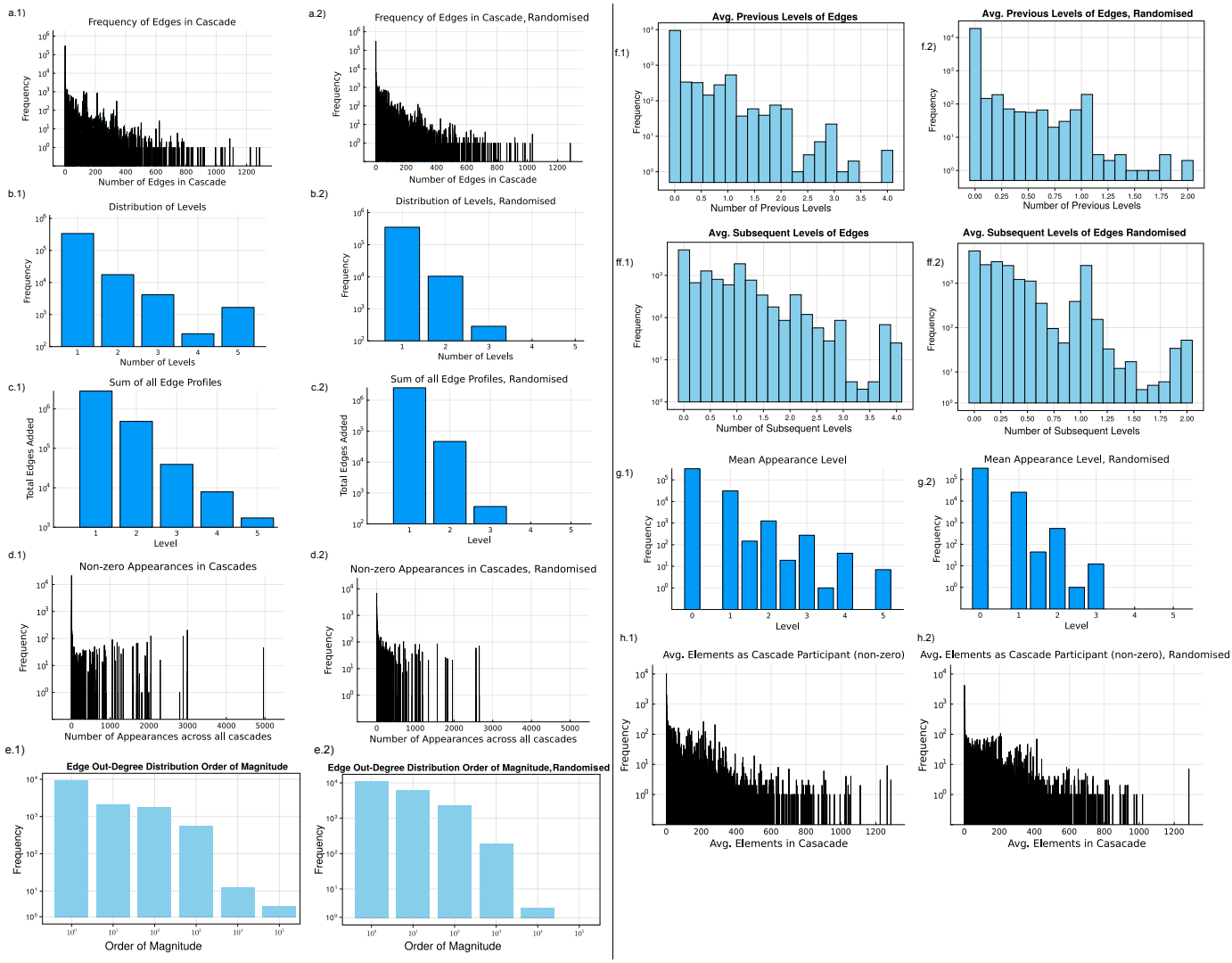

Figure 7.5: **Edge Metrics from Reactive Cascade: a.1)** Distribution of total elements in reactive cascades of all edges in the chemical space from the ORD. **a.2)** Dist. of total elements in reactive cascades in randomised null model hypergraph. **b.1)** Dist. of number of levels in reactive cascades of all edges in the chemical data. **b.2)** Dist. of number of levels in reactive cascades in the randomised null model. Notice that there are more levels in the data than in the randomisation. **c.1)** Cumulative Edge Profile of chemical data. **c.2)** Cumulative Edge Profile of randomisation. **d.1)** Frequency of edge appearances in cascades in chemical data. **d.2)** Freq. of edge appearances in cascades in randomisation. **e.1)** Edge Out-Degree dist. by order of magnitude in chemical data. **e.2)** Edge Out-Degree dist. in randomisation. **f.1)** Average previous level dist. across all edges in chemical data. **f.2)** Avg. previous level dist. across all edges in randomisation. **ff.1)** Avg. subsequent level dist. across all edges in chemical data. **ff.2)** Avg. subsequent level dist. across all edges in randomisation. **g.1)** Mean appearance level dist. across all edges in chemical reaction data. **g.2)** Mean appearance level dist. across all edges in randomisation. **h.1)** Dist. of avg. elements in cascade edge participates in for chemical data. **h.2)** Dist. of avg. elements in cascade edge participates in randomisation.

7.3.3.3 Node Metrics

Following the insights derived from edge metrics on reactive cascades of the chemical reaction data and the randomised null model, we examine the node metrics. In Figure 7.6, we demonstrate the application of the node metrics for reactive capital on the data recovered from the ORD (left hand side panels) and the randomised null model hypergraph (right hand side panels). Information from node metrics quantifies the reactive capital of nodes, demonstrating their capacity to be produced as well as their role in facilitating reactions across the chemical space. We again compare with the randomised null model in order to emphasise the interpretation of our results. The node metrics are presented in the order in which they are introduced in Section 7.2.

We notice first the total number of nodes in the dataset is on the order of magnitude of 10^5 while the total of number nodes that can be reached by a node considering all the edges it participates in is on the order of 10^3. This does not contradict the fact that any edge can reach is on the order of 10^4, as the addition of further edges is a recombination of available nodes. This order of magnitude discrepancy between the nodes that can be reached through the reactive cascade and the total nodes in the system emphasises the isolated nature of reactions within the system. Indeed, the total number of nodes that can be reached through a reactive cascade is 3.8% of all nodes in the chemical reaction dataset and 4.74% of the nodes in the randomised data.

Once again, we note that there are five levels in the reactive cascade in the chemical reaction data and three in the randomisation. In Figures 7.6a) we note distribution of unique nodes that can be reached from nodes in initial edges. We notice that many more unique nodes can be reached by the randomised data, emphasising that it is the chemical properties of the molecules themselves that give rise to the limitations in spreading in the chemical reaction data. We show the distribution of the average levels a node propagates into from an associated initial edge in Figure 7.6b). The more limited levels in the randomised data in combination with the wider element reach of nodes indicates a substantially different spreading process with respect to the chemical reaction data which is far more constrained and ordered. In Figures 7.6c), we show the total node sampling profile of all nodes. The sampling profile indicates at what level an available node is used as an input in, or in the tail of, a reaction. We notice that node sampling permeates into all levels in the chemical reaction data, however drops off rapidly in the randomised data. We attain addition information regarding sampling in Figures 7.6d) which show that most nodes have low weighted appearance scores, with smaller subsets of nodes having high weighted appearance

scores in the reaction data. In the randomised model, weighted appearance scores are more evenly distributed. These patterns demonstrate that the connectivity of the ORD leads to more in-depth traversal and has more node use selectivity than the randomised connectivity structure. In Figures 7.6e) we show the production appearances of nodes, that is, how many times nodes are produced, or appear as heads. This indicates that most molecules are produced few times, however there are subsets of nodes that experience thousands of productions. The upper bound of this metric is approximately twice as high in the data than are expected in the randomisation. This is demonstrative of chemical space properties. Firstly, there are chemical species that are commonly produced. This may be by importance of the chemical species or by their high association to other species. This property may also be introduced due to scientific exploration bias, that is, scientists exploring properties of familiar molecules or molecules that have important physical applications. This also indicates that there are molecules that have many routes of production, which of interest in the retrosynthesis process. In Figure 7.6f) we show the head out-degree of the two models. We note that the head out-degree contains an order of magnitude more in the metric measurement for the data than the randomisation. This emphasises that there is a small subset of highly influential molecules in the dataset. The more extended level structure of the chemical reaction data is also elucidated by this metric, as the level structure requires the addition of new heads which then allows for further traversal of the graph. The total number of nodes that have a non-zero head out-degree is 2.28% in the chemical reaction dataset and 3.37% in randomisation, supporting the interpretation that there is node selectivity due to chemical structure in the ORD. We also find that there is overlap between the nodes that contribute to production, as in have a non-zero head out-degree, and the nodes that can be reached from reactive cascades. Specifically, 59% of nodes that can be produced through reactions have a non-zero head out-degree and themselves contribute to further production. Figure 7.6g) and Figure 7.6h) show the distribution of the average number of previous and subsequent levels of all nodes in the cascade. Individually, this indicates the positioning of a node in the reactive cascades in which it participates in reference to the length of the cascades in which it is produced. This positioning highlights the accessibility of node production across the chemical hypergraph, for example, nodes that are often terminally positioned in relation to long cascades require more reactions to produce and these metrics reveal the average path dependence of the node. Figures 7.6i) show the mean first appearance of nodes in cascades, giving a reference for the level of production of the node. This emphasises

how quickly, or easily, a node can be produced, regardless of the length of the cascades in which it participates. This allows for a reference of reproducibility across cascades without the influence of individual cascade structures. Finally in Figures 7.6j) we show the distribution of the average size of cascades by elements in which nodes participate. We note that despite vastly different level structures, the average number of elements in these levels are remarkably similar in the chemical reaction data and the randomised data. For an individual molecule, the average size of elements in cascades in which the node participates will give its relative association to other molecules and production pathways in general.

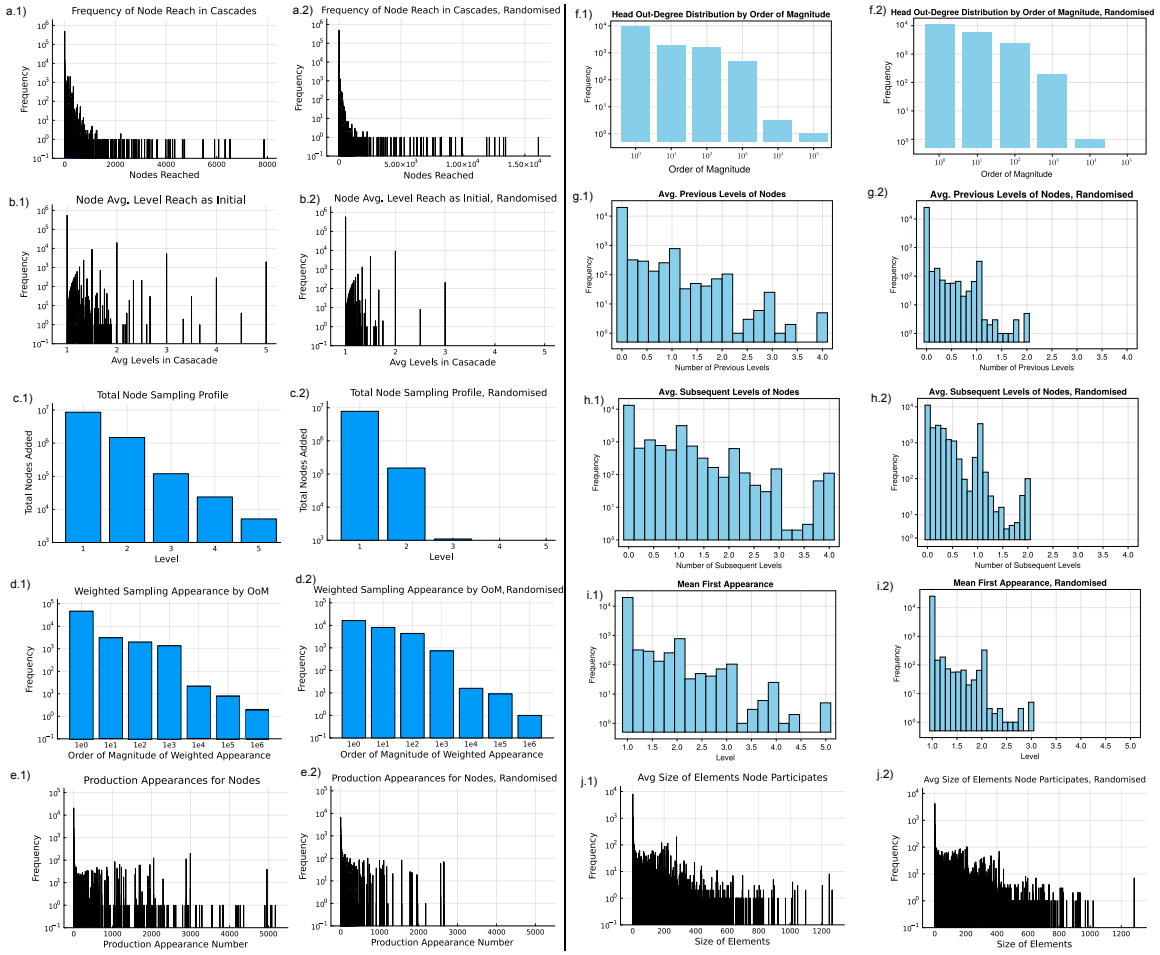

Figure 7.6: **Node Metrics from Reactive Cascade:**, **a.1** Distribution of unique nodes that can be reached by any node across initial edges in the chemical space from the ORD. **a.2** Dist. of unique nodes that can be reached by any node across initial edges in the randomised null model hypergraph.**b.1** Dist. of average levels nodes participating in an initial edge may reach across all reactive cascades in the chemical data. **b.2** Dist. of average levels nodes participating in an initial edge may reach across all reactive cascade in the randomised null model. **c.1** Dist. of Total Node Sampling Profile of all nodes in the chemical data. **c.2** Dist. of Total Node Sampling Profile of all nodes in the chemical space in the randomised null model. **d.1** Weighted sampling appearance of all nodes of chemical data by order of magnitude. **d.2** Weighted sampling appearance of all nodes of randomisation by order of magnitude. **e.1** Dist. of node production in the chemical data. **e.2** Dist. of node production in the randomisation. **f.1** Head Out-Degree dist. by order of magnitude in chemical data. **f.2** Head Out-Degree dist. by order of magnitude in randomisation. **g.1** Average previous level dist. of nodes in chemical data. **g.2** Avg. previous level dist. of nodes in randomisation. **h.1** Avg. subsequent level dist. of nodes in chemical data. **h.2** Avg. subsequent level dist. of nodes in randomisation. **i.1** Mean first appearance level dist. of nodes in chemical data. **i.2** Mean first appearance level dist. of nodes in randomisation. **j.1** Avg. node elements in cascades nodes participate in, dist. across all cascades in chemical data. **j.2** Avg. node elements in cascades nodes participate in, dist. across all cascades in randomisation.

7.3.4 Node Selection Reactive Cascade

We demonstrate the robustness of the reactive cascade and further insights into the chemical space by applying our methodology to a selected set of nodes as opposed to an initial edge. For this analysis, we consider two sets of nodes.

The first set of nodes we consider is the 10 nodes with the top reachability according to our previous reactive cascade application to the chemical reaction dataset. That is, the set of 10 nodes that when acting as initial nodes in any cascade, can reach the most unique other nodes across cascades. Applying the reactive cascade to this node set, we recover six levels. The cascade reaches 2,391 edges or reactions, and 2,336 unique new nodes or molecules. The selection of this set of nodes under the reactive cascade reaches more levels and reacts into vastly more edges than a reactive cascade starting from any one edge in the chemical reaction dataset.

The second set of nodes we consider is the 10 nodes with the top head out-degree according to the application of the reactive cascade to the chemical reaction dataset. Applying the reactive cascade to this node set, we recover 2 levels with 32 edges and 29 new unique nodes reached. The edge profiles of both selected node reactive cascade are visualised in Figure 7.7.

The first set considers nodes that forward react from all initial edges, or start points. The second set, on the other hand, considers the forward reaction of nodes that are made available by access through a reaction. A possible explanation of the differences in the amount of nodes and edges that can be reached between the two sets is that there are nodes in initial edges that have high potential for production, yet are not commonly produced. Therefore, these high production potential nodes will not be accounted for in the second set. Additionally, the nodes in the first set may more commonly appear together in initial edges that can reach many reactions. In the second set, the nodes with high head out-degree may not necessarily appear together such that there are other reactions whose tails are subsets of these nodes.

This analysis applies the reactive cascade to nodes of interest, demonstrating the flexibility of this framework for connectivity and graph exploration. By amending the starting set of nodes or edges, we can investigate different properties of the graph. The choice of starting set grants access to vastly different parts of the data and reveals different aspects of its organization. This node selection analysis highlights the properties of graph influence starting from different definitions of high impact molecules, showing that the reactive cascade methodology is amenable to different analytical goals. This analysis also further emphasises the importance of multiple properties that capture reactive capital.

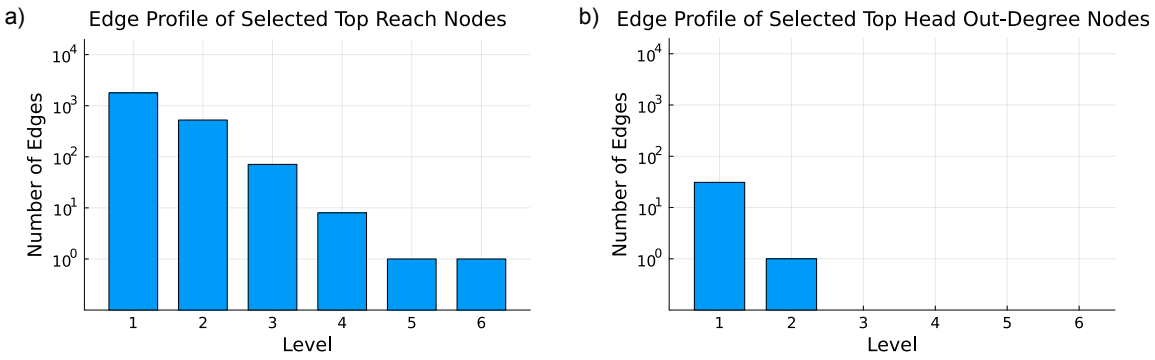

Figure 7.7: **Edge Profiles of Selected Node Cascades: a)** The edge profile of the reactive cascade beginning from the 10 top nodes of the dataset in edges that reach the most other nodes. **b)** The edge profile of the reactive cascade beginning from the 10 nodes in the dataset that have the highest head out-degree. Both plots are on a log-scale.

7.4 Discussion

In this chapter, we have developed a methodology of random walks on hypergraphs, which we term the *reactive cascade*. This methodology is inspired by chemical reactions, however, this random walk methodology is broadly applicable to systems governed by the existence and availability of precursors or prerequisites in order for percolation to occur. An example of such an application would be to the modelling of innovations as directed hypergraphs, where certain technologies are required to exist before the development of other advancements. The methodology could also be applied to the study of supply chains or tracking the development of medical techniques, where essential components or breakthroughs must be in place to proceed. Under the view of availability of nodes and hyperedges, the reactive cascade unveils system wide properties of connectivity and tracks sequential properties of the structure under consideration. Reactive cascades map the growth of possibilities and detail reachable states of systems beginning from determined seed points. These cascades give an idea of both the depth and breadth of walks unfolding from initial starts. The sequential ordering of reactive cascades unveils how fast and how far one can expand into the system space.

The reactive cascade metrics further explore the properties of connectivity and positioning of the species and interactions of the system. These metrics relate to the growth potential induced by species or interactions in a system. Edge metrics relate to the structure of interactions in the network and how these structures induce spreading behaviours. Node metrics quantify the production and use of system species

in spreading. Based on existing properties of the system, these metrics highlight structure and dynamics of the system based on a defined random walk. Reactive cascades reveal drivers of network growth and in particular the timing and sequential ordering of interacting factors. Importance of species based on user interest can be derived from reactive metrics.

We then applied our reactive cascade methodology to the chemical space, uncovering properties of the chemical space and also exploring the reactive capital of reactions and molecules, overall contributing to an understanding of their synthetic utility within a space of explored chemical reactions. This pertains specifically to a networks perspective of the ordering of the chemical space and the steps required to produce reactions or molecules, as well as their use in other reactions or molecules. We note that retrosynthetic paths are of interest to the scientific community and hypergraph models of chemical reactions are an active area of investigation. Our methodology and its application contribute to an embedding of desired retrosynthetic properties as a framework for similar retrosynthetic analyses.

Our results are multi-fold. We extract properties of the chemical space under the exploration of our reactive cascade methodology. We examine the reachability and connectivity of the chemical reaction space. In particular, under the consideration of single product reactions, we show that only a small subset of edges react into other edges. We also show overlap between chemical inputs and outputs, suggesting multiple production paths of molecules. Upon the analysis of the undirected connected component structure, we show the emergence of a large connected component, however, further results show a limited capacity of graph traversal from the selection of initial reactions when the direction of interactions is considered. The strong general associations of the space of all reactions and molecules and the small spreading capacity from selected edges emphasises the importance of the directionality in the chemical space. Specifically, intrinsic chemical properties induce a highly disconnected space under the reactive cascade exploration of availability.

We have also generated a null model randomisation of the chemical reaction data hypergraph and compare our developed edge and node metrics to the null model. These results allow us to emphasise the importance of ordering and layering of sequences of reactions in the chemical space. We compare the redundancy of the use and production of reactions in the chemical data to the randomisation, implying chemical properties of certain reactions and nodes introduce the specificity of further spreading on the chemical space structure. We show that some species have multiple

production paths, which may also be a property of the tendency of expertise with familiar molecules to explore the chemical space.

Our network based metrics identify the reactive capital of reactions and molecules within the chemical reaction data. We analyse the positioning and the potential of production of reactions and molecules. These can contribute to measures of their synthetic utility, as positioning in cascades may indicate more ease of synthesis as it may require more or less steps. We note that our measures are independent of molecular complexity estimates and instead focus on the network based role of edges and nodes to capture functional importance.

In conclusion, we develop a methodology of a random walk on directed hypergraphs. The application of our methodology to explore the chemical space unveils key features of its structure and interactions.

7.5 Conclusion and Further Directions

The development of the reactive cascade contributes to methods for the exploration of connectivity and structure of directed hypergraphs. In particular, the reactive cascade reveals structural properties as they relate to graph traversal. The reactive cascade also captures properties of both edges and nodes, allowing for analysis of a system at both scales of information. We may also curate the initial starting point of the methodology to pinpoint properties of interest, for example, by selecting a set of initial edges or nodes we wish to further investigate.

Further, edge and node metrics following the application of the reactive cascade reveal patterns of use and production of hyperedges and nodes. This allows us to quantify useful properties of their positioning, functionality, and influence. The potential for application of the developed methodology and associated metrics to diverse fields of investigation is strong. This approach may be used to explore the effects of combinatorial possibilities and sequential precedence in general. The applied analysis in Chapter 7 uses reactive cascades and their associated metrics to map the chemical space and explore the dynamical possibilities based on the structure of the chemical space. The results reveal the contribution of edges and molecules to the development of the chemical space. The application of the reactive cascade also identifies key reactions and molecules as well as highlights how intrinsic chemical properties lead to a sequential, multi-step connective structure. We further quantify pathways of interaction and demonstrate the limited potential for widespread graph traversal. This

investigation contributes to a framework for mapping and quantifying the chemical space and providing insight into its structure and potential for dynamical behaviour.

In order to further develop this analysis, we would require a precise dataset of chemical interactions without user mislabelling. In the current dataset, this would require the development of machine learning approaches to data processing. Other avenues of interest would be to include further information in order to quantify production capability, for example, production costs and machinery requirements or yields of molecules.

Chapter 8

Conclusion

Complexity emerges in systems across scales and fields, with a breadth of techniques and approaches applied to provide insight into such systems. Mathematical modelling techniques and methodological development are critical to investigating such systems and are essential to the interpretation of data in order to reveal system patterns, characterise system organisation, and understand emergent phenomena. In this thesis, we present investigations that develop insights into the dynamics and organisation of complex, interacting systems. Throughout our analysis, we introduce tools and methods that accompany data across scales and fields. These contributions are presented in this thesis in three parts.

In Part I, we develop a qualitative analysis of a cytokine signalling network. We investigate the robustness of an established network extraction methodology pertaining to sparse datasets. We subsequently analyse the effects of regulatory cytokines in the dynamical system imposed on the topology of the extracted cytokine network. Our analysis gives insight into the specific effects of cytokines on the production of IL-23, a cytokine of interest to the scientific community in the development of interventions in Inflammatory Bowel Disease. We extend the *in vitro* model to incorporate *in vivo* system features. Our results demonstrate the sensitivity of the system to cytokine inclusion or exclusion and emphasise the need for caution in translating model predictions between *in vitro* and *in vivo* models. These results have implications for scientists and modellers regarding the selection of which elements to incorporate into their experiments or models, as well as how readily their conclusions from *in vitro* models may be applicable to *in vivo* scenarios. In this study, we show how specific network nodes are influential across the network, emphasising how the connectivity and structure of the network is influential in the percolation of network dynamics. We show how the targeted modification of key cytokines alter the dynamics and production of cytokines throughout the network over time.

We note that the network extraction methodology and the subsequent analysis rely on a sparse dataset and a single measurement time. In future studies of the cytokine network system, the incorporation of additional measurement times may help capture further features of cytokine dynamics. Further, our study demonstrates shifts in cytokine concentrations following the inclusion and exclusion of certain cytokines. Hence, perturbation studies with an expanded number of cytokines may be insightful in further elucidating mechanisms for dynamical control of cytokines of interest.

In Part II, we identify patterns of data and information propagation online through mobile phone applications following catastrophic events. We identify features of applications that influence the patterns of their spread and give insight into human behavioural patterns as people navigate the online information space. We identify shifts in user behaviour in the short- and long- term following disruptive events. We also detail the effects of city-specific user behaviour on application traffic, including differences between data patterns in cities of incidence and in other cities.

In addition, we develop a methodology for demonstrating spatial propagation of online information and the creation of a null model for the comparison and tracking of information spread trends and their subsequent relaxation over time. We demonstrate that mobile phone application traffic, and correspondingly information dissemination, following an event shows a spatial organisation in a radial pattern emitting outward from the epicentre of the event. We subsequently demonstrate the relaxation of this radial pattern over time as a return to baseline activity. Using mobile phone application traffic data, we show the emergence of human behavioural trends and patterns in information propagation. This study contributes to insights of emergent collective behaviour following disruptive events.

Further avenues of research are plentiful. These include applications to different events, such as earthquakes, flooding, and mass casualty events. They may also be applied in general to disruptions beyond catastrophe, such as major sporting events or large-scale celebrations. An especially interesting direction is the the exploration large events in close proximity to each other at fine-grained spatial scale, in order to investigate potential intersections of radial spatial propagation of application traffic. Further, explorations of misinformation through application traffic is a promising avenue. For example, such an investigation could shed insight into whether misinformation spread is characterised by similar or distinct patterns as real world events. This information could also lead to a better understanding of how application features propagate misinformation. Such studies may inform the monitoring of applications

experiencing certain traffic patterns or suggest strategies to mitigate the spread of misinformation.

Finally, in Part III, we develop a methodology for revealing the organisation, connectivity, and propagation dynamics of directed hypergraphs. We subsequently develop metrics for ranking and investigating the influence and reach of hyperedges and nodes following the application of our novel methodology. This method contributes to the expanding framework of directed hypergraph-based analytic methodologies. The reactive cascade and associated metrics are readily applicable to further datasets and amenable to the selection of subsets of interest in directed hypergraphs for further investigation. In particular, these methods provide insight into directed hypergraphs that model systems requiring the satisfaction of certain sequential ordering combinations and greatly detail how influential certain subsets of these systems are.

The reactive cascade methodology we develop is inspired by reactivity in chemistry and we subsequently demonstrate how its application to a chemical dataset uncovers the structure and organisation of the chemical space. This contribution is particularly relevant given the growing interest in data-driven methodologies in chemistry [234]. We demonstrate the potential for propagation within the chemical space and elucidate properties of reactions and molecules within the chemical space. Further work in applications of this methodology to the chemical space could incorporate the inclusion of reaction rates, yields, and machinery requirements. In particular, this would be insightful in the retrosynthetic process of producing target molecules.

In this thesis, we explore complex systems across scales and fields, unveiling structures and developing novel relevant methodologies. Our investigations demonstrate the dynamics, propagation, and organisation of interacting systems and emphasise how these processes influence complex systems across multiple disciplines. Through the use of mathematical modelling, we identify features of complex systems and characterise emergent behaviours. Collectively, this thesis develops methodologies and contributes insights into data organisation, the propagation of dynamics, and the influence and patterns arising from various structures of complex systems.

A Appendix A

A.1 Constraints on network extraction due to the multi-start approach

We note that the methodology in [98] implements, among other methods, a multi-start approach with regards to minimisation of a log-likelihood function for network extraction. With the introduction of noise to the dataset, it is possible that the optimisation algorithm used in the modelling environment chooses initial starting points such that multiple local minima are found, creating an error in the network extraction pipeline. However, due to this stochastic sampling of the multi-start, a noisy simulation that encounters the error of multiple local minima can often be successfully run through the network extraction pipeline when the same noisy data-file used in the simulation is attempted multiple times. In the studies of Jansen et al. it was found that it might be necessary for other data sets to adjust the number of initial starting points, which are then sampled through Latin-hypercube sampling (LHS), and the acceptable parameter ranges, *i.e.* restricted lower and upper bounds for $\alpha_{u,i}$ and $\beta_{v,i}$ used to generate further starting points [98].

For further detail on the existence of multiple local minima, initial guess constraints, and parameter ranges, refer to [98], Appendix: Implementation and code availability, where the network extraction can be found.

A.2 Parameter values for regulatory cytokines

Model parameters describe the effect of one cytokine on another, or the effect of a stimulus on a cytokine. Parameters are given in the form, s_i, $\alpha_{u,i}$, and $\beta_{v,i}$. Parameters s_i indicate the effect of stimulation from LPS or IFNγ on the cytokine i. Parameters of the form $\alpha_{u,i}$ denote an up-regulatory effect of cytokine u on i. Parameters of the form $\beta_{v,i}$ denote a down-regulatory effect of cytokine v on i. The value of each parameter describes the strength of interaction, either from the stimulus or between cytokines.

Parameter values and the bounds of the logliklihood estimations of each parameter are given for $\alpha_{u,i}$ and $\beta_{v,i}$ for the regulatory cytokine network for the model equation (3.1) in Table 1. These bounds are produced by likelihood profiles estimates in the modelling environment, Data2Dynamics in MATLAB, in the parameter estimation process that is part of the network extraction [186, 185, 114]. The designations in

Table 1a are given to the cytokines for brevity.

Cytokine	Number label
IL-23	1
IL-10	2
IL-1β	3
IL-1α	4
IFNγ	5

(a)

Parameter	Parameter Value	Lower Bound	Upper Bound
s_1	0.0905	0.0252	0.1531
s_2	0.0755	0.0557	0.1013
s_3	0.2631	0.2343	0.2963
s_4	0.2184	0.1857	0.2693
α_{44}	0.065038	0.0207	0.1249
α_{43}	0.023289	0	0.0539
α_{42}	0.16651	0.0855	0.3101
α_{34}	0.19289	0.1340	0.2624
α_{33}	0.16976	0.1268	0.2175
α_{32}	0.54265	0.3648	0.7514
α_{31}	1.4733	0.6528	2.3097
α_{51}	3.4794	1.2668	6.5900
β_{24}	0.33024	0.2235	0.4707
β_{23}	0.20063	0.1380	0.2768
β_{22}	0.33443	0.1510	0.5592
β_{21}	22.8805	9.5702	35.6125

(b)

Table 1: (a) Cytokines and their corresponding number labels. These labels are used in equations of the form of equation (3.1) and in the parameters extracted from the network inference. (b) Parameter values for stimulus influence and up- and down-regulatory interaction strength between cytokines. These parameters are used in all equations of the form of equation (3.1). The range considered in the sensitivity analysis for each paramter is between the lower bound and upper bound given in the penultimate and final columns. Note that IFNγ does not have an s_i value (stimulation from LPS), as IFNγ is considered as a stimulatory agent.

B Appendix B: Part II

This appendix provides additional information for Part II of the report.

B.1 Full List of Applications Available in NetMob2023 Dataset

Table 2: Numbered list of applications and web services.

1. Amazon Web Services
2. Apple App Store
3. Apple Mail
4. Apple Music
5. Apple Siri
6. Apple Video
7. Apple Web Services
8. Apple iCloud
9. Apple iMessage
10. Apple iTunes
11. Clash of Clans
12. DailyMotion
13. Deezer
14. Dropbox
15. EA Games
16. Facebook
17. Facebook Live
18. Facebook Messenger
19. Fortnite
20. Google Docs
21. Google Drive
22. Google Mail
23. Google Maps
24. Google Meet
25. Facebook Messenger
26. Fortnite
27. Google Docs
28. Google Drive
29. Google Mail
30. Google Maps
31. Google Meet
32. Google Play Store
33. Google Web Services
34. Instagram
35. LinkedIn
36. Microsoft Azure
37. Microsoft Mail
38. Microsoft Office
39. Microsoft Skydrive
40. Microsoft Store
41. Microsoft Web Services
42. Molotov
43. Netflix
44. Orange TV
45. Periscope
46. Pinterest
47. PlayStation
48. Pokemon GO
49. Web Downloads
50. Web Finance
51. Web Food
52. Web Games
53. Web Streaming
54. Web Transportation
55. Web Weather
56. Web e-Commerce
57. WhatsApp
58. Wikipedia
59. Yahoo
60. Yahoo Mail
61. YouTube

Bibliography

[1] Nicole Abaid and Maurizio Porfiri. Fish in a ring: spatio-temporal pattern formation in one-dimensional animal groups. *Journal of The Royal Society Interface*, 7(51):1441–1453, April 2010. Publisher: Royal Society.

[2] Allison E. Aiello, Audrey Renson, and Paul N. Zivich. Social Media– and Internet-Based Disease Surveillance for Public Health. *Annual Review of Public Health*, 41(1):101–118, April 2020. Publisher: Annual Reviews.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974. Conference Name: IEEE Transactions on Automatic Control.

[4] Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer New York, New York, NY, 1998. Series Title: Springer Series in Statistics.

[5] Leman Akoglu and Christos Faloutsos. Event detection in time series of mobile communication graphs. In *Proceedings of the Army Science Conference*, pages 77–79, 2010.

[6] Sinan G. Aksoy, Cliff Joslyn, Carlos Ortiz Marrero, Brenda Praggastis, and Emilie Purvine. Hypernetwork science via high-order hypergraph walks. *EPJ Data Science*, 9(1):1–34, December 2020. Number: 1 Publisher: SpringerOpen.

[7] Shahriar Akter and Samuel Fosso Wamba. Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, 26(2):173–194, May 2016.

[8] Shayan Alipour, Alessandro Galeazzi, Emanuele Sangiorgio, Michele Avalle, Ljubisa Bojic, Matteo Cinelli, and Walter Quattrociocchi. Cross-platform social dynamics: an analysis of ChatGPT and COVID-19 vaccine conversations.

Scientific Reports, 14(1):2789, February 2024. Publisher: Nature Publishing Group.

[9] Akira Andoh, Yuhki Yagi, Makoto Shioya, Atsushi Nishida, Tomoyuki Tsujikawa, and Yoshihide Fujiyama. Mucosal cytokine network in inflammatory bowel disease. *World Journal of Gastroenterology : WJG*, 14(33):5154–5161, September 2008.

[10] Ana Andres-Hernando, Belda Dursun, Christopher Altmann, Nilesh Ahuja, Zhibin He, Rhea Bhargava, Charles E. Edelstein, Alkesh Jani, Thomas S. Hoke, Christina Klein, and Sarah Faubel. Cytokine production increases and cytokine clearance decreases in mice with bilateral nephrectomy. *Nephrology Dialysis Transplantation*, 27(12):4339–4347, December 2012.

[11] Muhammad Anshari, Mohammad Nabil Almunawar, Syamimi Ariff Lim, and Abdullah Al-Mudimigh. Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics*, 15(2):94–101, July 2019.

[12] Dominik Aschenbrenner, Maria Quaranta, Soumya Banerjee, Nicholas Ilott, Joanneke Jansen, Boyd Steere, Yin-Huai Chen, Stephen Ho, Karen Cox, Carolina V. Arancibia-Cárcamo, Mark Coles, Eamonn Gaffney, Simon PL Travis, Lee Denson, Subra Kugathasan, Jochen Schmitz, Fiona Powrie, Stephen N. Sansom, and Holm H. Uhlig. Deconvolution of monocyte responses in inflammatory bowel disease reveals an IL-1 cytokine network that regulates IL-23 in genetic and acquired IL-10 resistance. *Gut*, 70(6):1023–1036, June 2021. Publisher: BMJ Publishing Group Section: Inflammatory bowel disease.

[13] Tom S. Au, Rong Duan, Heeyoung Kim, and Guang-Qin Ma. Spatiotemporal Event Detection in Mobility Network. In *2010 IEEE International Conference on Data Mining*, pages 28–37, December 2010. ISSN: 2374-8486.

[14] James P. Bagrow, Dashun Wang, and Albert-László Barabási. Collective response of human populations to large-scale emergencies. *PLOS ONE*, 6(3), March 2011.

[15] Ruth Baker. Further mathematical biology lecture notes. Lecture notes, 2023.

[16] Philip Ball. Should biology put complexity first? *Cell Systems*, 16(2):101197, February 2025.

[17] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, March 2018.

[18] Hugo Barbosa, Fernando B. de Lima-Neto, Alexandre Evsukoff, and Ronaldo Menezes. The effect of recency to human mobility. *EPJ Data Science*, 4(1):21, December 2015.

[19] René Barone and Michel Chanon. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *Journal of Chemical Information and Computer Sciences*, 41(2):269–272, March 2001. Publisher: American Chemical Society.

[20] Federico Battiston, Enrico Amico, Alain Barrat, Ginestra Bianconi, Guilherme Ferraz de Arruda, Benedetta Franceschiello, Iacopo Iacopini, Sonia Kéfi, Vito Latora, Yamir Moreno, Micah M. Murray, Tiago P. Peixoto, Francesco Vaccarino, and Giovanni Petri. The physics of higher-order interactions in complex systems. *Nature Physics*, 17(10):1093–1098, October 2021. Publisher: Nature Publishing Group.

[21] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports*, 874:1–92, August 2020.

[22] Linus Bengtsson, Jean Gaudart, Xin Lu, Sandra Moore, Erik Wetter, Kankoe Sallah, Stanislas Rebaudet, and Renaud Piarroux. Using Mobile Phone Data to Predict the Spatial Spread of Cholera. *Scientific Reports*, 5(1):8923, March 2015. Publisher: Nature Publishing Group.

[23] Linus Bengtsson, Xin Lu, Anna Thorson, Richard Garfield, and Johan von Schreeb. Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. *PLOS Medicine*, 8(8):e1001083, August 2011. Publisher: Public Library of Science.

[24] Katrin Bennhold and James Glanz. Notre-Dame's Safety Planners Underestimated the Risk, With Devastating Results, April 2019. `https://www.nytimes.com/2019/04/19/world/europe/notre-dame-fire-safety.html`.

[25] Steven H. Bertz. Convergence, molecular complexity, and synthetic analysis. *Journal of the American Chemical Society*, 104(21):5801–5803, October 1982.

[26] Purnima Bholowalia and Arvind Kumar. EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, 105(9), 2014.

[27] Christian Bick, Elizabeth Gross, Heather A. Harrington, and Michael T. Schaub. What are higher-order networks? *SIAM Review*, 65(3):686–731, August 2023. arXiv:2104.11329 [cs].

[28] Billeto. Concert Ed Sheeran Groupama Stadium Lyon 24 mai 2019. Event listing on Billetto, May 2019. `https://billetto.fr/en/e/concert-ed-sheeran-groupama-stadium-lyon-24-mai-2019-billets-362084`.

[29] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175–308, February 2006.

[30] U. Boehm, T. Klamp, M. Groot, and J. C. Howard. Cellular responses to interferon-gamma. *Annual Review of Immunology*, 15:749–795, 1997. PMID: 9143706.

[31] Jean Pierre Boon, James F. Lutsko, and Christopher Lutsko. Microscopic approach to nonlinear reaction-diffusion: The case of morphogen gradient formation. *Physical Review E*, 85(2):021126, February 2012. Publisher: American Physical Society.

[32] Guillaume Bouguen. Recent advances in cytokines: Therapeutic implications for inflammatory bowel diseases. *World Journal of Gastroenterology*, 17(5):547, 2011.

[33] Alain Bretto. Hypergraphs: Basic concepts. In *Hypergraph Theory*, Mathematical Engineering, pages 1–21. Springer International Publishing AG, Switzerland, 2013.

[34] Jonathan Briskman. Top Apps Worldwide for Q1 2019 by Downloads, May 2019.

[35] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, January 2006. Publisher: Nature Publishing Group.

[36] Gordon Broderick, Jim Fuite, Andrea Kreitz, Suzanne D. Vernon, Nancy Klimas, and Mary Ann Fletcher. A formal analysis of cytokine networks in Chronic Fatigue Syndrome. *Brain, Behavior, and Immunity*, 24(7):1209–1217, October 2010.

[37] Thomas Böttcher. An Additive Definition of Molecular Complexity. *Journal of Chemical Information and Modeling*, 56(3):462–470, March 2016. Publisher: American Chemical Society.

[38] R Callard, AJT George, and J Stark. Cytokines, chaos, and complexity. *Immunity*, 11(5):507–513, November 1999.

[39] Timoteo Carletti, Federico Battiston, Giulia Cencetti, and Duccio Fanelli. Random walks on hypergraphs. *Physical Review E*, 101(2), February 2020. arXiv:1911.06523 [physics].

[40] Flávia Castro, Ana Patrícia Cardoso, Raquel Madeira Gonçalves, Karine Serre, and Maria José Oliveira. Interferon-Gamma at the Crossroads of Tumor Immune Surveillance or Evasion. *Frontiers in Immunology*, 9, 2018.

[41] Joseph E. Cavanaugh and Andrew A. Neath. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, 11(3):e1460, 2019.

[42] Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.

[43] N. Chatzistefani, M. J. Chappell, C. Hutchinson, S. Kletzenbauer, and N. D. Evans. A mathematical model characterising Achilles tendon dynamics in flexion. *Mathematical Biosciences*, 284:92–102, February 2017.

[44] Bernard Chazelle. The Convergence of Bird Flocking. *Journal of the ACM*, 61(4):1–35, July 2014.

[45] Ziqi Chen, Oluwatosin R. Ayinde, James R. Fuchs, Huan Sun, and Xia Ning. G2retro as a two-step graph generative models for retrosynthesis prediction. *Communications Chemistry*, 6(1), May 2023.

[46] Connor W. Coley, Luke Rogers, William H. Green, and Klavs F. Jensen. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *Journal of*

Chemical Information and Modeling, 58(2):252–261, February 2018. Publisher: American Chemical Society.

[47] Jean Frédéric Colombel, Paul Rutgeerts, Walter Reinisch, Dirk Esser, Yanxin Wang, Yinghua Lang, Colleen W. Marano, Richard Strauss, Björn J. Oddens, Brian G. Feagan, Stephen B. Hanauer, Gary R. Lichtenstein, Daniel Present, Bruce E. Sands, and William J. Sandborn. Early Mucosal Healing With Infliximab Is Associated With Improved Long-term Clinical Outcomes in Ulcerative Colitis. *Gastroenterology*, 141(4):1194–1201, October 2011. Publisher: Elsevier.

[48] Jean Frédéric Colombel, William J. Sandborn, Walter Reinisch, Gerassimos J. Mantzaris, Asher Kornbluth, Daniel Rachmilewitz, Simon Lichtiger, Geert D'Haens, Robert H. Diamond, Delma L. Broussard, Kezhen L. Tang, C. Janneke van der Woude, and Paul Rutgeerts. Infliximab, Azathioprine, or Combination Therapy for Crohn's Disease. *New England Journal of Medicine*, 362(15):1383–1395, April 2010. Publisher: Massachusetts Medical Society.

[49] T. M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006.

[50] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, October 2008. arXiv:0803.2189 [physics].

[51] Mengyao Cui. Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. *Geoscience and Remote Sensing*, 3(1):9–16, 2020.

[52] A. C. Davison. *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK, 1st edition, 2003.

[53] Guilherme Ferraz de Arruda, Giovanni Petri, and Yamir Moreno. Social contagion models on hypergraphs. *Physical Review Research*, 2(2):023032, April 2020. Publisher: American Physical Society.

[54] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

[55] Charles A. Dinarello. Historical Review of Cytokines. *European Journal of Immunology*, 37(Suppl 1):S34–S45, 2007.

[56] Christl Donnelly. Prelims statistics and data analysis lectures 11–16. Lecture notes, University of Oxford, 2025.

[57] J. L. Doob. Probability and Statistics. *Transactions of the American Mathematical Society*, 36(4):759–775, 1934. Publisher: American Mathematical Society.

[58] Aurélien Ducournau and Alain Bretto. Random walks in directed hypergraphs and application to semi-supervised image segmentation. *Computer Vision and Image Understanding*, 120:91–102, March 2014.

[59] Nathan Eagle, Alex Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278, 2009.

[60] Justin Eilertsen, Santiago Schnell, and Sebastian Walcher. The unreasonable effectiveness of the total quasi-steady state approximation, and its limitations. *Journal of Theoretical Biology*, 583:111770, April 2024.

[61] Joshua M. Epstein and Robert L. Axtell. *Growing Artificial Societies: Social Science from the Bottom Up.* The MIT Press, October 1996.

[62] Radek Erban, S Jonathan Chapman, and Philip K Maini. A practical guide to stochastic simulations of reaction–diffusion processes. *Diffusion Processes*, 2007.

[63] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):1–11, December 2009. Publisher: BioMed Central.

[64] Ernesto Estrada. What is a complex system, after all? *Foundations of Science*, 29(4):1143–1170, December 2024.

[65] Stephen Eubank, Hasan Guclu, V. S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, Zoltán Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, May 2004. Publisher: Nature Publishing Group.

[66] Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Cluster Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, West Sussex, UK, 5th edition, 2011.

[67] Mauro Faccin. Measuring dynamical systems on directed hyper-graphs. *Physical Review E*, 106(3):034306, September 2022. arXiv:2202.12810 [physics].

[68] Deema Ibrahim Fallatah and Hafeez Aderinsayo Adekola. Digital epidemiology: harnessing big data for early detection and monitoring of viral outbreaks. *Infection Prevention in Practice*, 6(3):100382, September 2024.

[69] Fei Fang and Spencer P. Lake. Modelling approaches for evaluating multiscale tendon mechanics. *Interface Focus*, 6(1):20150044, February 2016. Publisher: Royal Society.

[70] Soheil Feizi, Daniel Marbach, Muriel Médard, and Manolis Kellis. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology*, 31(8):726–733, August 2013.

[71] Marcin Fialkowski, Kyle J. M. Bishop, Victor A. Chubukov, Christopher J. Campbell, and Bartosz A. Grzybowski. Architecture and Evolution of Organic Chemistry. *Angewandte Chemie International Edition*, 44(44):7263–7269, 2005.

[72] Matthias Friedrich, Mathilde Pohin, and Fiona Powrie. Cytokine Networks in the Pathophysiology of Inflammatory Bowel Disease. *Immunity*, 50(4):992–1006, April 2019. Publisher: Elsevier.

[73] DM Frucht, T Fukao, C Bogdan, H Schindler, JJ O'Shea, and S Koyasu. IFN-gamma production by antigen-presenting cells: mechanisms emerge. *Trends in Immunology*, 22(10):556–560, October 2001.

[74] Angel Garcia-Chung, Marisol Bermúdez-Montaña, Peter F. Stadler, Jürgen Jost, and Guillermo Restrepo. Chemically inspired Erdős–Rényi hypergraphs. *Journal of Mathematical Chemistry*, 62(6):1357–1383, July 2024.

[75] Ruth Garcia-Gavilanes, Milena Tsvetkova, and Taha Yasseri. Dynamics and biases of online attention: the case of aircraft crashes. *Royal Society Open Science*, 3(10), October 2016.

[76] Mihai Georgescu, Nattiya Kanhabua, Daniel Krause, Wolfgang Nejdl, and Stefan Siersdorfer. Extracting Event-Related Information from Article Updates in Wikipedia. In *Advances in Information Retrieval*, volume 7814 of *Lecture Notes in Computer Science*, pages 254–266, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[77] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, December 1977.

[78] Valerie J. Gillet, Glenn Myatt, Zsolt Zsoldos, and A. Peter Johnson. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspectives in Drug Discovery and Design*, 3(1):34–50, December 1995. Company: Springer Distributor: Springer Institution: Springer Label: Springer Publisher: Kluwer Academic Publishers.

[79] J. P. Gisbert, A. C. Marín, A. G. McNicholl, and M. Chaparro. Systematic review with meta-analysis: the efficacy of a second anti-TNF in patients with inflammatory bowel disease whose previous anti-TNF treatment has failed. *Alimentary Pharmacology & Therapeutics*, 41(7):613–623, April 2015.

[80] Shayane Martins Rodrigues Gomes, Andréia Carolinne de Souza Brito, Wânia Ferraz Pereira Manfro, Marcelo Ribeiro-Alves, Roberto Stefan de Almeida Ribeiro, Mariana Soares da Cal, Vinicius da Cunha Lisboa, Daniel Paiva Barros de Abreu, Leda dos Reis Castilho, Luís Cristóvão de Moares Sobrino Porto, Thiago Thomáz Mafort, Agnaldo José Lopes, Silvia Amaral Gonçalves da Silva, Patrícia Maria Lourenço Dutra, and Luciana Silva Rodrigues. High levels of pro-inflammatory SARS-CoV-2-specific biomarkers revealed by in vitro whole blood cytokine release assay (CRA) in recovered and long-COVID-19 patients. *PLOS ONE*, 18(4):e0283983, April 2023.

[81] Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008. Publisher: Nature Publishing Group.

[82] Leo A. Goodman. Snowball Sampling. *The Annals of Mathematical Statistics*, 32(1):148–170, 1961. Publisher: Institute of Mathematical Statistics.

[83] Paul A. Gore. 11 - Cluster Analysis. In *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, pages 297–321. Academic Press, San Diego, January 2000.

[84] Neta Gotlieb, Ella Rosenne, Pini Matzner, Lee Shaashua, Liat Sorski, and Shamgar Ben-Eliyahu. The misleading nature of in vitro and ex-vivo findings in studying the impact of stress hormones on NK cell cytotoxicity. *Brain, behavior, and immunity*, 45:277–286, March 2015.

[85] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

[86] Ibtissem Guenfoud and Edith Honan. French launch manhunt for suspect behind lyon blast that injured 13. ABC News, May 2019.

[87] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd edition, 2012.

[88] Stephen B Hanauer, Brian G Feagan, Gary R Lichtenstein, Lloyd F Mayer, S Schreiber, Jean Frederic Colombel, Daniel Rachmilewitz, Douglas C Wolf, Allan Olson, Weihang Bao, and Paul Rutgeerts. Maintenance infliximab for Crohn's disease: the ACCENT I randomised trial. *The Lancet*, 359(9317):1541–1549, May 2002.

[89] Xingsheng He and Yu-Ru Lin. Measuring and monitoring collective attention during shocking events. *EPJ Data Science*, 6(1):30, December 2017.

[90] John R. Hershey and Peder A. Olsen. Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–317–IV–320, April 2007. ISSN: 2379-190X.

[91] Lori Hinnant. 66 minutes: The frantic race to save Notre Dame. Associated Press, April 2019.

[92] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, August 1952.

[93] Jun Hu, Yuxin Wang, and Ping Li. Online city-scale hyper-local event detection via analysis of social media and human mobility. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 626–635, December 2017.

[94] Qi Huang, Lin-Li Li, and Sheng-Yong Yang. RASA: A Rapid Retrosynthesis-Based Scoring Method for the Assessment of Synthetic Accessibility of Drug-like Molecules. *Journal of Chemical Information and Modeling*, 51(10):2768–2777, October 2011. Publisher: American Chemical Society.

[95] Heinrich J. Huber and Hitesh B. Mistry. Explaining in-vitro to in-vivo efficacy correlations in oncology pre-clinical development via a semi-mechanistic mathematical model. *Journal of Pharmacokinetics and Pharmacodynamics*, 51(2):169–185, April 2024.

[96] Shoichi Ishida, Kei Terayama, Ryosuke Kojima, Kiyosei Takasu, and Yasushi Okuno. AI-Driven Synthetic Route Design Incorporated with Retrosynthesis Knowledge. *Journal of Chemical Information and Modeling*, 62(6):1357–1367, March 2022. Publisher: American Chemical Society.

[97] J. E. Jansen. *Mathematical modelling of macrophage functioning in inflammatory bowel disease*. Dphil thesis, University of Oxford, 2020. Supervisors: E. Gaffney, H. Uhlig, and M. Coles; deposited in ORA on 23 October 2022.

[98] Joanneke E. Jansen, Dominik Aschenbrenner, Holm H. Uhlig, Mark C. Coles, and Eamonn A. Gaffney. A method for the inference of cytokine interaction networks. *PLOS Computational Biology*, 18(6):e1010112, June 2022.

[99] Thomas Jun Jewell, Andrew L. Krause, Philip K. Maini, and Eamonn A. Gaffney. Chase-and-Run and Chirality in Nonlocal Models of Pattern Formation, May 2025. arXiv:2505.17372 [nlin].

[100] A. Peter Johnson, Chris Marshall, and Philip N. Judson. Starting material oriented retrosynthetic analysis in the LHASA Program. 1. General Description. *Journal of Chemical Information and Computer Sciences*, 32(5):411–417, September 1992.

[101] Jürgen Jost and Raffaella Mulas. Hypergraph laplace operators for chemical reaction networks. *Advances in Mathematics*, 351:870–896, July 2019.

[102] Jeremy Kahn. Facebook Activates 'Safety Check' Amid Brussels Attack Crisis. Bloomberg News, online, March 2016.

[103] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine Piatko, Ruth Silverman, and Angela Y. Wu. The analysis of a simple k -means clustering

algorithm. In *Proceedings of the sixteenth annual symposium on Computational geometry*, pages 100–109, Clear Water Bay Kowloon Hong Kong, May 2000. ACM.

[104] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, October 2008. Publisher: Nature Publishing Group.

[105] Steven M. Kearnes, Michael R. Maser, Michael Wleklinski, Anton Kast, Abigail G. Doyle, Spencer D. Dreher, Joel M. Hawkins, Klavs F. Jensen, and Connor W. Coley. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021. PMID: 34727496.

[106] James Keener and James Sneyd. *Mathematical Physiology: II: Systems Physiology*, volume 8/2 of *Interdisciplinary Applied Mathematics*. Springer New York, New York, NY, 2nd edition, 2009.

[107] David Kiron, Pamela Kirk Prentice, and Renee Boucher Ferguson. The analytics mandate. MIT Sloan Management Review Project, May 2014.

[108] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and Cellular Networks. *PLOS Computational Biology*, 5(5):e1000385, May 2009. Publisher: Public Library of Science.

[109] Leopold Klein and Dietmar W. Hutmacher. Straddling the Line Between In Vitro and Ex Vivo Investigations. *Tissue Engineering. Part C, Methods*, 30(10):443–451, October 2024.

[110] Zoe Kleinman. Google launches new 'AI mode' search feature in UK. BBC News, July 2025. Section: Technology.

[111] Ryota Kobayashi, Patrick Gildersleve, Takeaki Uno, and Renaud Lambiotte. Modeling Collective Anticipation and Response on Wikipedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):315–326, May 2021.

[112] B. Kolic. *Data-driven methods for social complex systems: Social media, data assimilation and how they connect*. Dphil thesis, University of Oxford, 2022. Supervisor: J. Doyne Farmer; also supervised by C. Hepburn; Mathematics Institute, MPLS Division.

[113] Petra Krause, Venetia Morris, Jason A. Greenbaum, Yoon Park, Unni Bjoerheden, Zbigniew Mikulski, Tracy Muffley, Jr-Wen Shui, Gisen Kim, Hilde Cheroutre, Yun-Cai Liu, Bjoern Peters, Mitchell Kronenberg, and Masako Murai. IL-10-producing intestinal macrophages prevent excessive antibacterial innate immunity by limiting IL-23 synthesis. *Nature Communications*, 6(1):7055, May 2015.

[114] Clemens Kreutz, Andreas Raue, and Jens Timmer. Likelihood based observability analysis and confidence intervals for predictions of dynamic models. *BMC Systems Biology*, 6(1):120, September 2012.

[115] A. Kuceyeski, S. Shah, J. P. Dyke, S. Bickel, F. Abdelnour, N. D. Schiff, H. U. Voss, and A. Raj. The application of a mathematical model linking structural and functional connectomes in severe brain injury. *NeuroImage: Clinical*, 11:635–647, January 2016.

[116] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. Publisher: Institute of Mathematical Statistics.

[117] James Ladyman, James Lambert, and Karoline Wiesner. What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67, January 2013.

[118] Renaud Lambiotte. Networks course notes, 2022. Lecture notes, University of Oxford.

[119] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes. From networks to optimal higher-order models of complex systems. *Nature Physics*, 15(4):313–320, April 2019. Publisher: Nature Publishing Group.

[120] Renaud Lambiotte and Michael T. Schaub. *Modularity and Dynamics on Complex Networks*. Elements in the Structure and Dynamics of Complex Networks. Cambridge University Press, 2022.

[121] Patricia Lamirande, Eamonn A. Gaffney, Michael Gertz, Philip K. Maini, Jessica R. Crawshaw, and Antonello Caruso. A First-Passage Model of Intravitreal Drug Delivery and Residence Time—Influence of Ocular Geometry, Individual Variability, and Injection Location. *Investigative Ophthalmology & Visual Science*, 65(12):21, October 2024.

[122] Gregory Landrum. RDKit: Open-source cheminformatics software. `https://www.rdkit.org`. Accessed: 2025.

[123] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, December 2008.

[124] Wilmer Leal and Guillermo Restrepo. Formal structure of periodic system of elements. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2224), April 2019.

[125] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 251–260, New York, NY, USA, 2012. Association for Computing Machinery.

[126] Chun-Qing Li and Wei Yang. 2 - Essential reliability methods. In *Time-Dependent Reliability Theory and Its Applications*, Woodhead Publishing Series in Civil and Structural Engineering, pages 51–119. Woodhead Publishing, January 2023.

[127] Andreas A. Linninger, Michalis Xenos, Brian Sweetman, Sukruti Ponkshe, Xiaodong Guo, and Richard Penn. A mathematical model of blood, cerebrospinal fluid and brain dynamics. *Journal of Mathematical Biology*, 59(6):729–759, December 2009.

[128] Chao Liu, Dewei Chu, Kourosh Kalantar-Zadeh, Jacob George, Howard A. Young, and Guozhen Liu. Cytokines: From Clinical Significance to Quantification. *Advanced Science*, 8(15):2004433, 2021.

[129] Yu Liu, Cole Mathis, Michał Dariusz Bajczyk, Stuart M. Marshall, Liam Wilbraham, and Leroy Cronin. Exploring and mapping chemical space with molecular assembly trees. *Science Advances*, 7(39):eabj2465, 2021.

[130] Eugenio Llanos, Wilmer Leal, Andrés Bernal, Jürgen Jost, and Peter F. Stadler. Are the chemical families still there? Formal structure of similarity of elements and its thermochemical domain. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 480(2298):20240165, September 2024. Publisher: Royal Society.

[131] Eugenio J. Llanos, Wilmer Leal, Duc H. Luu, Jürgen Jost, Peter F. Stadler, and Guillermo Restrepo. Exploration of the chemical space and its three historical regimes. *Proceedings of the National Academy of Sciences of the United States of America*, 116(26):12660–12665, June 2019.

[132] J. David Logan. *Applied Mathematics*. John Wiley & Sons, Hoboken, NJ, 4th edition, 2013.

[133] Philipp Lorenz-Spreen, Bjarke Mørch Mønsted, Philipp Hövel, and Sune Lehmann. Accelerating dynamics of collective attention. *Nature Communications*, 10(1):1759, 2019.

[134] Thomas Louail, Maxime Lenormand, Oliva G Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific reports*, 4(1):5276, 2014.

[135] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[136] Yonghui Ma. The Story of "Story" — A Simple Timeline, August 2020.

[137] Franco Maceri, Michele Marino, and Giuseppe Vairo. An insight on multiscale tendon modeling in muscle–tendon integrated behavior. *Biomechanics and Modeling in Mechanobiology*, 11(3):505–517, March 2012.

[138] James B MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297. University of California Press, 1967.

[139] Philip K. Maini, Thomas E. Woolley, Ruth E. Baker, Eamonn A. Gaffney, and S. Seirin Lee. Turing's model for biological pattern formation and the robustness problem. *Interface Focus*, 2(4):487–496, August 2012.

[140] Kevin J. Maloy and Fiona Powrie. Intestinal homeostasis and its breakdown in inflammatory bowel disease. *Nature*, 474(7351):298–306, June 2011.

[141] Vipul Mann and Venkat Venkatasubramanian. AI-driven hypergraph network of organic chemistry: network statistics and applications in reaction classification. *Reaction Chemistry & Engineering*, 8(3):619–635, February 2023. Publisher: The Royal Society of Chemistry.

[142] Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(1):S7, March 2006.

[143] Simeone Marino, Ian B. Hogue, Christian J. Ray, and Denise E. Kirschner. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology*, 254(1):178–196, September 2008.

[144] Orlando E Martínez-Durive, Sachit Mishra, Cezary Ziemlicki, Stefania Rubrichi, Zbigniew Smoreda, and Marco Fiore. The netmob23 dataset: A high-resolution multi-region service-level mobile data traffic cartography. *arXiv preprint arXiv:2305.06933*, 2023.

[145] Oliver Mason and Mark Verwoerd. Graph theory and networks in biology. *IET Systems Biology*, 1(2):89–119, 2007.

[146] Hisatake Matsumoto, Hiroshi Ogura, Kentaro Shimizu, Mitsunori Ikeda, Tomoya Hirose, Hiroshi Matsuura, Sujin Kang, Kanae Takahashi, Toshio Tanaka, and Takeshi Shimazu. The clinical importance of a cytokine network in the acute phase of sepsis. *Scientific Reports*, 8:13995, September 2018.

[147] Giovanni Mauro, Nicola Pedreschi, Renaud Lambiotte, and Luca Pappalardo. Dynamic models of gentrification. *Advances in Complex Systems*, 28(06):2540006, July 2025.

[148] Sofia Medina, Shazia'Ayn Babul, Timothy Larock, Rohit Sahasrabuddhe, Renaud Lambiotte, and Nicola Pedreschi. Detection of anomalous spatio-temporal patterns of app traffic in response to catastrophic events. *EPJ Data Science*, 14(1), May 2025.

[149] Meta. Introducing Instagram Reels. Instagram Blog, August 2020.

[150] Meta Platforms, Inc. Crisis Response. https://www.facebook.com/about/crisisresponse/v2, 2020.

[151] P. J. Mitchell and J. D. Murray. Facilitated diffusion: The problem of boundary conditions. *Biophysik*, 9(3):177–190, September 1973.

[152] Anja Mittag and Attila Tárnok. Basics of standardization and calibration in cytometry – a review. *Journal of Biophotonics*, 2(8-9):470–481, 2009.

[153] James Morris. Notre Dame fire: Hundreds of mourners gather in paris for emotional vigil, April 2019. Accessed: 2024-07-16.

[154] Kevin A. Murgas, Emil Saucan, and Romeil Sandhu. Hypergraph geometry reflects higher-order dynamics in protein interaction networks. *Scientific Reports*, 12(1):20879, December 2022. Publisher: Nature Publishing Group.

[155] James D Murray. Mathematical biology second, corrected edition. In *Mathematical Biology*, volume 19. Springer Berlin / Heidelberg, Germany, 1993.

[156] Elon Musk. Grok 3 launch announcement. X (formerly Twitter) post, August 2025. "Grok 3 ... the 'smartest AI on Earth.' Release scheduled for Monday at 8 PM PT" (World Government Summit, Dubai); accessed 20 August 2025.

[157] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100, February 2003.

[158] Oscar Méndez-Lucio and José L. Medina-Franco. The many roles of molecular complexity in drug discovery. *Drug Discovery Today*, 22(1):120–126, January 2017.

[159] Ramón Nartallo-Kaluarachchi, Renaud Lambiotte, and Alain Goriely. From reductionism to realism: Holistic mathematical modelling for complex biological systems, March 2025. arXiv:2503.20511 [physics].

[160] Markus F. Neurath. Current and emerging therapeutic targets for IBD. *Nature Reviews Gastroenterology & Hepatology*, 14(5):269–278, May 2017.

[161] M. E. J. Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

[162] M. E. J. Newman. Complex Systems: A Survey. *American Journal of Physics*, 79(8):800–810, August 2011. arXiv:1112.1440 [cond-mat].

[163] Mark Newman. *Networks: An Introduction*. Oxford University Press, March 2010.

[164] BBC News. TikTok named as the most downloaded app of 2020. *BBC News*, August 2021.

[165] NHS Transformation Directorate. NHS COVID-19 app. `https://transform.england.nhs.uk/covid-19-response/nhs-covid-19-app/`, 2020.

[166] Steven A. Niederer, Kenneth S. Campbell, and Stuart G. Campbell. A short history of the development of mathematical models of cardiac mechanics. *Journal of Molecular and Cellular Cardiology*, 127:11–19, February 2019.

[167] Denis Noble. The development of mathematical models of the heart. *Chaos, Solitons & Fractals*, 5(3):321–333, March 1995.

[168] Nuria Oliver, Bruno Lepri, Harald Sterly, Renaud Lambiotte, Sébastien Deletaille, Marco De Nadai, Emmanuel Letouzé, Albert Ali Salah, Richard Benjamins, Ciro Cattuto, Vittoria Colizza, Nicolas de Cordes, Samuel P. Fraiberger, Till Koebe, Sune Lehmann, Juan Murillo, Alex Pentland, Phuong N. Pham, Frédéric Pivetta, Jari Saramäki, Samuel V. Scarpino, Michele Tizzoni, Stefaan Verhulst, and Patrick Vinck. Mobile phone data for informing public health actions across the covid-19 pandemic life cycle, 2020.

[169] Tudor I. Oprea and Cristian Bologa. Molecular Complexity: You Know It When You See It. *Journal of Medicinal Chemistry*, 66(18):12710–12714, September 2023.

[170] Miles Osborne, Saša Petrovic, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Bieber no more: First story detection using twitter and wikipedia. In *Sigir 2012 workshop on time-aware information access*, pages 16–76, 2012.

[171] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1998.

[172] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Community dynamics in social networks. In *Noise and Stochastics in Complex Systems and Finance*, volume 6601, page 660106. International Society for Optics and Photonics, SPIE, 2007.

[173] Joao Palotti, Natalia Adler, Alfredo Morales-Guzman, Jeffrey Villaveces, Vedran Sekara, Manuel Garcia Herranz, Musa Al-Asad, and Ingmar Weber. Monitoring of the Venezuelan exodus through Facebook's advertising platform. *PLOS ONE*, 15(2):e0229175, February 2020. Publisher: Public Library of Science.

[174] Luca Pappalardo, Salvatore Rinzivillo, and Filippo Simini. Human Mobility Modelling: Exploration and Preferential Return Meet the Gravity Model. *Procedia Computer Science*, 83:934–939, January 2016.

[175] Kari Paul. Instagram Reels: Facebook's history of 'adapting' competitors' familiar features. The Guardian, aug 2020.

[176] Elian Peltier, James Glanz, Mika Gröndahl, Weiyi Cai, Adam Nossiter, and Liz Alderman. Notre-Dame came far closer to collapsing than people knew. This is how it was saved. *The New York Times*, July 2019. `https://www.nytimes.com/interactive/2019/07/16/world/europe/notre-dame.html`.

[177] Peter Pivonka and Colin R Dunstan. Role of mathematical modeling in bone fracture healing. *BoneKEy Reports*, 1:221, November 2012.

[178] Peter Pivonka and Svetlana V. Komarova. Mathematical modeling in bone biology: From intracellular signaling to tissue mechanics. *Bone*, 47(2):181–189, August 2010.

[179] Spyros E. Polykalas and George N. Prezerakos. When the mobile app is free, the product is your personal data. *Digital Policy, Regulation and Governance*, 21(2):89–101, March 2019. Publisher: Emerald.

[180] Giulia Preti, Adriano Fazzone, Giovanni Petri, and Gianmarco De Francisci Morales. Higher-order null models as a lens for social systems. *Physical Review X*, 14(3), August 2024.

[181] Daniele Pretolani. Finding hypernetworks in directed hypergraphs. *European Journal of Operational Research*, 230(2):226–230, October 2013.

[182] Stephen R. Proulx, Daniel E. L. Promislow, and Patrick C. Phillips. Network thinking in ecology and evolution. *Trends in Ecology & Evolution*, 20(6):345–353, June 2005.

[183] Gonzalo Marcelo Ramírez-Ávila, Jürgen Kurths, Stéphanie Depickère, and Jean-Louis Deneubourg. Modeling Fireflies Synchronization. In *A Mathematical Modeling Approach from Nonlinear Dynamics to Complex Systems*, pages 131–156. Springer International Publishing, Cham, 2019.

[184] Milan Randić and Dejan Plavšić. Characterization of molecular complexity. *International Journal of Quantum Chemistry*, 91(1):20–31, 2003.

[185] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, August 2009.

[186] A. Raue, B. Steiert, M. Schelker, C. Kreutz, T. Maiwald, H. Hass, J. Vanlier, C. Tönsing, L. Adlung, R. Engesser, W. Mader, T. Heinemann, J. Hasenauer, M. Schilling, T. Höfer, E. Klipp, F. Theis, U. Klingmüller, B. Schöberl, and J. Timmer. Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, 31(21):3558–3560, November 2015.

[187] Guillermo Restrepo. Chemical space: limits, evolution and modelling of an object bigger than our universal library. *Digital Discovery*, 1(5):568–585, 2022. Publisher: Royal Society of Chemistry.

[188] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong. On the Levy-Walk Nature of Human Mobility. *IEEE/ACM Transactions on Networking*, 19(3):630–643, June 2011.

[189] Giulia Roda, Bindia Jharap, Narula Neeraj, and Jean-Frederic Colombel. Loss of Response to Anti-TNFs: Definition, Epidemiology, and Management. *Clinical and Translational Gastroenterology*, 7:e135, January 2016.

[190] A. Saltelli, S. Tarantola, and K. P.-S. Chan. A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output. *Technometrics*, 41(1):39–56, February 1999.

[191] Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. *Sensitivity analysis in practice: a guide to assessing scientific models.* John Wiley & Sons, Ltd, Chichester, UK, 2004.

[192] J. A. Sanz-Herrera, J. M. Garcia-Aznar, and M. Doblare. A mathematical model for bone tissue regeneration inside a specific type of scaffold. *Biomechanics and Modeling in Mechanobiology*, 7(5):355–366, October 2008.

[193] N. Scafetta. The complex planetary synchronization structure of the solar system. *Pattern Recognition in Physics*, 2(1):1–19, January 2014.

[194] Thomas C. Schelling. Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2):143–186, July 1971.

[195] S Schnell. Enzyme Kinetics at High Enzyme Concentration. *Bulletin of Mathematical Biology*, 62(3):483–499, May 2000.

[196] Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H. Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science*, 11(12):3316–3325, March 2020. Publisher: The Royal Society of Chemistry.

[197] Lee A. Segel and Marshall Slemrod. The Quasi-Steady-State Assumption: A Case Study in Perturbation. *SIAM Review*, 31(3):446–477, 1989. Publisher: Society for Industrial and Applied Mathematics.

[198] Shrabani Sen, Syed Shahed Riaz, and Deb Shankar Ray. Temperature dependence and temperature compensation of kinetics of chemical oscillations; Belousov–Zhabotinskii reaction, glycolysis and circadian rhythms. *Journal of Theoretical Biology*, 250(1):103–112, January 2008.

[199] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27, 1948.

[200] Julie Shapiro. Facebook Activates Safety Check for Orlando Nightclub Shooting. TIME, June 2016.

[201] Abhishek Sharma, Dániel Czégel, Michael Lachmann, Christopher P. Kempes, Sara I. Walker, and Leroy Cronin. Assembly theory explains and quantifies selection and evolution. *Nature*, 622(7982):321–328, October 2023. Publisher: Nature Publishing Group.

[202] Shi Shen, Nikita Murzintcev, Changqing Song, and Changxiu Cheng. Information retrieval of a disaster event from cross-platform social media. *Information Discovery and Delivery*, 45(4):220–226, January 2017. Publisher: Emerald Publishing Limited.

[203] M.S.R. Shoaib, M.A. Haque, and Md. Asaduzzaman. Mathematical modeling of the heart. In *International Conference on Electrical & Computer Engineering (ICECE 2010)*, pages 626–629, December 2010.

[204] Parisa Shooshtari, Edgardo S. Fortuno III, Darren Blimkie, Miao Yu, Arvind Gupta, Tobias R. Kollmann, and Ryan R. Brinkman. Correlation analysis of intracellular and secreted cytokines via the generalized integrated mean fluorescence intensity. *Cytometry Part A*, 77A(9):873–880, 2010.

[205] Harsh Shrivastava, Xiuwei Zhang, Le Song, and Srinivas Aluru. GRNUlar: A Deep Learning Framework for Recovering Single-Cell Gene Regulatory Networks. *Journal of Computational Biology*, 29(1):27–44, January 2022. Publisher: Mary Ann Liebert, Inc., publishers.

[206] Alexander F. Siegenfeld and Yaneer Bar-Yam. An Introduction to Complex Systems Science and its Applications. *Complexity*, 2020:1–16, July 2020. arXiv:1912.05088 [physics].

[207] Filippo Simini, Marta C. González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, April 2012. Publisher: Nature Publishing Group.

[208] Matthew J. Simpson and Ruth E. Baker. Corrected mean-field models for spatially dependent advection-diffusion-reaction phenomena. *Physical Review E*, 83(5):051922, May 2011. Publisher: American Physical Society.

[209] Sitabhra Sinha. From network structure to dynamics and back again: Relating dynamical stability and connection topology in biological complex systems. *arXiv preprint*, arXiv:0804.0977, 2008. 15 pages, to appear in Proceedings of "Dynamics On and Of Complex Networks", ECSS'07 Satellite Workshop.

[210] Jürgen Sladeczek, Alexander J. Hartemink, and Joshua Robinson. *Banjo User Guide*. Duke University.

[211] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, October 2010. Publisher: Nature Publishing Group.

[212] Chenxiao Song and Reiichiro Kawai. Monte Carlo and variance reduction methods for structural reliability analysis: A comprehensive review. *Probabilistic Engineering Mechanics*, 73:103479, July 2023.

[213] Guillaume St-Onge, Iacopo Iacopini, Vito Latora, Alain Barrat, Giovanni Petri, Antoine Allard, and Laurent Hébert-Dufresne. Influential groups for seeding and sustaining nonlinear contagion in heterogeneous hypergraphs. *Communications Physics*, 5(1):25, January 2022. Publisher: Nature Publishing Group.

[214] M A Syakur, B K Khotimah, E M S Rochman, and B D Satoto. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, 336(1):012017, April 2018. Publisher: IOP Publishing.

[215] Jacqueline M. Tarrant. Blood Cytokines as Biomarkers of In Vivo Toxicity in Preclinical Safety Assessment: Considerations for Their Use. *Toxicological Sciences*, 117(1):4–16, September 2010.

[216] Mohammadamin Tavakoli, Alexander Shmakov, Francesco Ceccarelli, and Pierre Baldi. Rxn Hypergraph: a Hypergraph Attention Model for Chemical Reaction Representation, January 2022. arXiv:2201.01196 [physics].

[217] Giuseppe Tenti, Siv Sivaloganathan, and James M. Drake. Mathematical modeling of the brain: principles and challenges. *Neurosurgery*, 62(5):1146, May 2008.

[218] The Economist. Building with big data: The data revolution is changing the landscape of business. `http://www.economist.com/node/18741392`, May 2011.

[219] Stefan Thurner, Rudolf Hanel, and Peter Klimekl. *Introduction to the Theory of Complex Systems*. Oxford University Press, September 2018.

[220] Trang P. Tran. Personalized ads on Facebook: An effective marketing tool for online marketers. *Journal of Retailing and Consumer Services*, 39:230–242, November 2017.

[221] Holm H. Uhlig and Fiona Powrie. Translating Immunology into Therapeutic Concepts for Inflammatory Bowel Disease. *Annual Review of Immunology*, 36(1):755–781, April 2018.

[222] Vaiva Vasiliauskaite and Fernando E. Rosas. Understanding complexity via network theory: a gentle introduction, April 2020. arXiv:2004.14845 [physics].

[223] José Miguel Vela, Rafael Maldonado, Michel Hamon, Raimund Mannhold, Hugo Kubinyi, and Gerd Folkers. *In Vivo Models for Drug Discovery*. John Wiley & Sons, Incorporated, Newark, GERMANY, 2014.

[224] Jan Vilček and Marc Feldmann. Historical review: Cytokines as therapeutics and targets of therapeutics. *Trends in Pharmacological Sciences*, 25(4):201–209, April 2004.

[225] Valérie Voorsluijs, Ioannis G. Kevrekidis, and Yannick De Decker. Nonlinear behavior and fluctuation-induced dynamics in the photosensitive Belousov–Zhabotinsky reaction. *Physical Chemistry Chemical Physics*, 19(33):22528–22537, 2017. Publisher: Royal Society of Chemistry.

[226] P. Waage and C. M. Gulberg. Studies concerning affinity. *Journal of Chemical Education*, 63(12):1044, December 1986.

[227] Wavestone (formerly NewVantage Partners). Data and AI leadership executive survey, 2024. Technical Report 12, Wavestone, January 2024.

[228] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, February 1988.

[229] Jonathan P. Whiteley, David J. Gavaghan, and Clive E. W. Hahn. Mathematical modelling of oxygen transport to tissue. *Journal of Mathematical Biology*, 44(6):503–522, June 2002.

[230] George M. Whitesides and Rustem F. Ismagilov. Complexity in Chemistry. *Science*, 284(5411):89–92, April 1999. Publisher: American Association for the Advancement of Science.

[231] H. W. Whitlock. On the Structure of Total Synthesis of Complex Natural Products. *The Journal of Organic Chemistry*, 63(22):7982–7989, October 1998. Publisher: American Chemical Society.

[232] Daniel S. Wigh, Joe Arrowsmith, Alexander Pomberger, Kobi C. Felton, and Alexei A. Lapkin. Orderly: Datasets and benchmarks for chemical reaction data. *Journal of Chemical Information and Modeling*, 64(9):3790–3798, 2024.

[233] S.S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

[234] Wendy L. Williams, Lingyu Zeng, Tobias Gensch, Matthew S. Sigman, Abigail G. Doyle, and Eric V. Anslyn. The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS Central Science*, 7(10):1622–1637, October 2021. Publisher: American Chemical Society.

[235] Brandon A. Wright and Richmond Sarpong. Molecular complexity as a driving force for the advancement of organic synthesis. *Nature reviews. Chemistry*, 8(10):776–792, October 2024.

[236] Fang Wu and Bernardo A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

[237] Naijia Xiao, Aifen Zhou, Megan L. Kempher, Benjamin Y. Zhou, Zhou Jason Shi, Mengting Yuan, Xue Guo, Linwei Wu, Daliang Ning, Joy Van Nostrand, Mary K. Firestone, and Jizhong Zhou. Disentangling direct from indirect relationships in association networks. *Proceedings of the National Academy of Sciences*, 119(2):e2109995119, January 2022. Publisher: Proceedings of the National Academy of Sciences.

[238] Yang Xu, Paolo Santi, and Carlo Ratti. Beyond distance decay: Discover homophily in spatially embedded social networks. *Annals of the American Association of Geographers*, 112(2):505–521, 2022.

[239] Yining Xu, Neha Shrestha, Véronique Préat, and Ana Beloqui. An overview of *in vitro*, *ex vivo* and *in vivo* models for studying the transport of drugs across intestinal barriers. *Advanced Drug Delivery Reviews*, 175:113795, August 2021.

[240] Takahiro Yabe, Nicholas K. W. Jones, P. Suresh C. Rao, Marta C. Gonzalez, and Satish V. Ukkusuri. Mobile phone location data for disasters: A review from natural hazards and epidemics. *Computers, Environment and Urban Systems*, 94:101777, June 2022.

[241] Jyoti Yadav and Monika Sharma. A Review of K-mean Algorithm. *International Journal of Engineering Trends and Technology*, 4(7), 2013.

[242] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, page 177–186, New York, NY, USA, 2011. Association for Computing Machinery.

[243] Jiue-An Yang, Ming-Hsiang Tsou, Krzysztof Janowicz, Keith C Clarke, and Piotr Jankowski. Reshaping the urban hierarchy: patterns of information diffusion on social media. *Geo-spatial Information Science*, 22(3):149–165, 2019.

[244] Wayne W. Zachary. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

[245] Juan Zhao, Yiwei Zhou, Xiujun Zhang, and Luonan Chen. Part mutual information for quantifying direct associations in networks. *Proceedings of the National Academy of Sciences*, 113(18):5130–5135, May 2016. Publisher: Proceedings of the National Academy of Sciences.

[246] Marinka Zitnik, Michelle M Li, Aydin Wells, Kimberly Glass, Deisy Morselli Gysi, Arjun Krishnan, T M Murali, Predrag Radivojac, Sushmita Roy, Anaïs Baudot, Serdar Bozdag, Danny Z Chen, Lenore Cowen, Kapil Devkota, Anthony Gitter, Sara J C Gosline, Pengfei Gu, Pietro H Guzzi, Heng Huang, Meng Jiang, Ziynet Nesibe Kesimoglu, Mehmet Koyuturk, Jian Ma, Alexander R Pico, Nataša Pržulj, Teresa M Przytycka, Benjamin J Raphael, Anna Ritz, Roded Sharan, Yang Shen, Mona Singh, Donna K Slonim, Hanghang Tong, Xinan Holly Yang, Byung-Jun Yoon, Haiyuan Yu, and Tijana Milenković. Current and future directions in network biology. *Bioinformatics Advances*, 4(1):vbae099, 2024.

[247] Zuckerberg, Mark. This morning we activated safety check for people affected by the earthquake in nepal. Facebook post (on Mark Zuckerberg's timeline), April 2015.