Deep Graphs

Represent and Analyze Heterogeneous Complex Systems across Scales

DISSERTATION

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)
im Fach Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
Dipl.-Phys. Dominik Traxl

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. habil. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:

1. Jürgen Kurths
2. Manoel Cardoso
3. Bernd Blasius

Tag der mündlichen Prüfung: 5. Mai 2017

to my mother

Abstract

The tremendous amount of data that is generated and accessible these days (often referred to as "Big Data") provides a great opportunity from a scientific perspective. However, it also poses methodological challenges, particularly since more and more of that information originates in unstructured form. To analyze data, methods from traditional disciplines such as probability theory, multivariate statistics and non-linear dynamics are employed, but also new tools are being developed, especially in the fields of machine learning and deep learning. Regarding the representation of systems, network theory has proven to be a powerful instrument. Yet, even in its latest and most general form (i.e., multilayer networks), it is still lacking essential qualities to serve as a general data analysis framework. These include, most importantly, an explicit association of information with the nodes and edges of a network, and a conclusive representation of groups of nodes and their respective interrelations on different scales. I consider these qualities to be crucial in the representation of systems, but also in their analysis. First, because they facilitate the means to represent features and relations on different scales, and second, because they allow us to coarse-grain, simplify and highlight important large-scale structures in a data-driven analysis. The implementation of these qualities into a generalized framework is the primary contribution of this dissertation. I develop a network-based framework capable of representing heterogeneous complex systems across scales. As opposed to other frameworks in the scientific literature, groups of objects (supernodes) and their respective interrelations (superedges) are incorporated into a self-contained network representation. Furthermore, potentially unstructured and diverse information is explicitly associated with the different (super)nodes and (super)edges. For these reasons, my framework is capable of acting as a go-between, joining a unified and generalized network representation of systems with the statistical tools of traditional fields, as well as the methods developed in the rising field of machine learning. In combination with the software package that accompanies this dissertation, my framework, *deep graphs*, thus makes an important contribution to the field of complex systems and potentially to data analytics in general.

A number of applications of my framework are demonstrated. By constructing a deep graph of extreme rainfall events, I conduct an explorative analysis of spatio-temporal rainfall clusters and find propagation patterns that have not yet been identified in the meteorological literature. Based on the constructed deep graph, I provide statistical evidence that the spatio-temporally integrated size distribution of extreme rainfall clusters does not - as previously suggested - follow a powerlaw. Instead, I find that the size distribution over the oceans is best approximated by an exponentially truncated powerlaw. By means of a generative storm-track model, I argue that the exponential truncation of the observed distribution could be caused by the presence of land masses. In another application, I combine two high-resolution satellite products to identify spatio-temporal clusters of fire-affected areas in the Brazilian Amazon and characterize their land use specific burning conditions. Finally, I investigate the effects of white noise and global coupling strength on the maximum degree of synchronization for a variety of oscillator models coupled according to a broad spectrum of network topologies. I find a general sigmoidal scaling and validate it with a suitable regression model.

Zusammenfassung

Die enorme Menge an Daten die heutzutage zur Verfügung steht - und mit “Big Data” einen Sammelbegriff gefunden hat - bietet ungeahnte Möglichkeiten für die Wissenschaft. Allerdings gehen mit der Datenmenge auch methodische Herausforderungen einher, vor allem wenn man bedenkt, dass ein großer Teil der Daten die heute erzeugt werden unstrukturiert und heterogen sind. Zur Analyse von Daten werden Methoden aus traditionellen Fachgebieten, wie z.B. der Wahrscheinlichkeitsrechnung, der multivariaten Statistik und der nicht-linearen Dynamik herangezogen. Aber auch neue Methoden werden entwickelt, speziell in den Gebieten des maschinellen Lernens und des ”Deep Learnings”. Zur Darstellung von Systemen hat sich die Theorie von Netzwerken als besonders zweckdienlich herausgestellt. Jedoch fehlen in der Netzwerkdarstellung von Systemen - selbst in ihrer jüngsten und allgemeinsten Abwandlung (sogenannte ”multilayer networks”) - noch immer essentielle Bausteine um diese generell zur Datenanalyse heranzuziehen zu können. Allen voran fehlt es an einer expliziten Assoziation von Informationen mit den Knoten und Kanten eines Netzwerks und einer schlüssigen Darstellung von Gruppen von Knoten und deren Relationen auf verschiedenen Skalen. Diese Bausteine halte ich für besonders wichtig, sowohl in der Darstellung von komplexen Systemen, als auch für deren Analyse. Erstens erlauben sie die Darstellung von Eigenschaften und Relationen auf verschiedenen Skalen eines Systems, und zweitens ermöglichen sie eine granulare Vereinfachung von Systemen um besonders wichtige, großskalige Eigenschaften herauszustellen. Das Hauptaugenmerk dieser Dissertation ist der Einbindung dieser Bausteine in eine verallgemeinerte Rahmenstruktur gewidmet. Ich entwickle eine Netzwerkbasierte Rahmenstruktur die es ermöglicht heterogene, komplexe Systeme über sämtliche Skalen hinweg zu repräsentieren. Im Gegensatz zu anderen Rahmenstrukturen in der wissenschaftlichen Literatur integriere ich Gruppen von Knoten (Superknoten) und deren Relationen (Superkanten) in eine in sich geschlossene Netzwerkdarstellung. Außerdem werden potentiell unstrukturierte und vielfältige Information explizit mit den (Super)Knoten und (Super)Kanten des Netzwerks assoziiert. Aus diesen Gründen ist meine Rahmenstruktur in der Lage als Bindeglied zwischen einer vereinheitlichten und generalisierten Netzwerkdarstellung von Systemen, den statistischen Methoden traditioneller Fachgebiete, sowie den Methoden des maschinellen Lernens zu fungieren. In Verbindung mit dem von mir entwickelten Softwarepaket stellt diese verallgemeinerte Netzwerkdarstellung, *Deep Graphs*, einen wichtigen Beitrag zur Theorie der komplexen Systeme, und wahrscheinlich zur Datenanalyse im generellen, dar.

Eine Reihe von Anwendungen meiner Rahmenstruktur werden ebenfalls dargestellt. Ich konstruiere einen Deep Graph von globalen, extremen Regenereignissen zur explorativen Analyse von raumzeitlich ausgedehnten Regenfallclustern, und finde dabei Ausbreitungsmuster die in der wissenschaftlichen Literatur noch nicht identifiziert wurden. Auf Grundlage des Regenfall Deep Graphs liefere ich einen statistischen Beleg, dass die raumzeitlich integrierte Größenverteilung von Extremregenfallclustern keinem Potenzgesetz folgt, wie in der Literatur vermutet wurde. Stattdessen zeige ich, dass die Größenverteilung der Extremregenfallcluster über den Ozeanen am besten durch ein exponentiell gedämpftes Potenzgesetz beschrieben wird. Mit Hilfe eines generativen Sturm-Modells zeige ich, dass die exponentielle Dämpfung der beobachteten Größenverteilung durch das Vorhandensein von Landmasse auf unserem Planeten zustande kommen könnte. In einer weiteren Anwendung meiner Netzwerkdarstellung verknüpfe ich zwei hochauflösende Satelliten-Produkte um raumzeitliche Cluster von Feuer-betroffenen Gebieten im brasilianischen Amazonas zu identifizieren, und deren Landnutzungsspezifischen Brandeigenschaften zu charakterisieren. Zuletzt untersuche ich den Einfluss von weißem Rauschen und der globalen Kopplungsstärke auf die maximale Synchronisierbarkeit von Oszillatoren-Netzwerken für eine Vielzahl von Oszillatoren-Modellen, welche durch ein breites Spektrum an Netzwerktopologien gekoppelt sind. Ich finde ein allgemeingültiges sigmoidales Skalierungsverhalten, und validiere dieses mit einem geeignetem Regressionsmodell.

List of publications

This dissertation is partly based on the following publications. The identifiers given below (e.g. P1) are cited in the text to highlight passages that are connected to one or more of these papers.

P1 **D. Traxl**, N. Boers, J. Kurths, *General scaling of maximum degree of synchronization in noisy complex networks*, New Journal of Physics **16**, 115009 (2014).

P2 **D. Traxl**, N. Boers, J. Kurths, *Deep graphs - A general framework to represent and analyze heterogeneous complex systems across scales*, Chaos **26(6)**, 065303 (2016).

P3 **D. Traxl**, N. Boers, A. Rheinwalt, B. Goswami, J. Kurths, *The size distribution of spatiotemporal extreme rainfall clusters around the globe*, Geophysical Research Letters (accepted).

P4 **D. Traxl**, A. Cano-Crespo, M. Cardoso, K. Thonicke, J. Kurths, *A statistical characterization of fire-cluster burning conditions in the Brazilian Amazon,* in preparation.

Dominik Traxl

Berlin, May 16, 2017

Acknowledgements

This work was mainly developed within the scope of the IRTG 1740 / TRP 2011/50151–0, funded by the DFG / FAPESP. Part of the research was funded by the BMBF (FKZ 01GQ1001A).

Contents

List of Figures

List of Tables

List of frequently used mathematical symbols

G	a general, directed graph, given by a pair $G = (V, E)$
V	the set of nodes of graph G
E	the set of edges of graph G
G^p	a supergraph of G induced by a partition function p
V^p	the set of supernodes of supergraph G^p
E^p	the set of superedges of supergraph G^p
$G^{\underline{p}}$	a supergraph of G induced by an intersection partition function $\underline{p}$
$V^{\underline{p}}$	the set of supernodes of supergraph $G^{\underline{p}}$
$E^{\underline{p}}$	the set of superedges of supergraph $G^{\underline{p}}$
n	the number of nodes of graph G
n^p	the number of supernodes of supergraph G^p
$n^{\underline{p}}$	the number of supernodes of graph $G^{\underline{p}}$
$n^{p,i}$	the number of nodes in supernode i
$n^{\underline{p},\underline{i}}$	the number of nodes in supernode $\underline{i}$
$n^{\underline{p},\underline{i}}_{\text{types}}$	the number of types of features in supernode $\underline{i}$
$n^{\underline{p},\underline{i}}_t$	the number of features of type t in supernode $\underline{i}$
${}^{\underline{p}}n^{\underline{p},\underline{i}}_{\text{types}}$	the number of partition-specific types of features in supernode $\underline{i}$
f_i	the number of features in node i
n_{types}	the number of distinct types of features in G
n_t	the number of features of type t in G
m	the number of edges of graph G
m^p	the number of superedges of supergraph G^p
$m^{\underline{p}}$	the number of superedges of supergraph $G^{\underline{p}}$
$m^{p,ij}$	the number of edges in superedge (i, j)
$m^{\underline{p},\underline{i}\underline{j},\underline{r}}$	the number of edges in superedge $(\underline{i}, \underline{j}, \underline{r})$
$m^{\underline{p},\underline{i}\underline{j},\underline{r}}_{\text{types}}$	the number of types of relations in superedge $(\underline{i}, \underline{j}, \underline{r})$
$m^{\underline{p},\underline{i}\underline{j},\underline{r}}_t$	the number of relations of type t in supernode $(\underline{i}, \underline{j}, \underline{r})$
${}^{\underline{p}}m^{\underline{p},\underline{i}\underline{j},\underline{r}}_{\text{types}}$	the number of partition-specific types of relations in superedge $(\underline{i}, \underline{j}, \underline{r})$
r_{ij}	the number of relations in edge (i, j)
m_{types}	the number of distinct types of relations in G
m_t	the number of relations of type t in G

Chapter 1.

Introduction

At the present time, we are observing a quantification of our world at an unprecedented rate. On the one hand – due to the rapid technological progress – we are extracting an ever increasing amount of information from nature, ranging from subatomic to astronomical scales. On the other hand, we are producing a vast amount of information in our daily lives interacting with electronic devices, thereby generating traceable information, tracked and stored by us personally, but also by organizations, companies and governments.

From a scientific point of view, this rapid increase in the amount and heterogeneity of available data poses both a great opportunity, but also methodological challenges: how can we describe and represent complex systems, made of multifarious subsystems interacting intricately on various scales; and once we have a suitable representation, how do we detect structures, patterns and correlations therein, develop and test hypotheses and eventually come up with models and working theories of underlying mechanisms?

Rich tool sets to tackle these questions have been developed in the past, such as probability theory (Jaynes, 2005), multivariate statistics (Anderson, 2003), nonlinear dynamics (Strogatz, 1994; Thiel et al., 2010) and game theory (Osborne and Rubinstein, 1994). Additionally, new methodologies to deal with the immense amount of information are being developed, especially in the fields of machine learning and deep learning (Hastie et al., 2009; Bishop, 2006; Haykin, 2009; Vapnik, 1998; Deng and Yu, 2014).

When it comes to the representation of data, network theory - which models the relations between a system's constituent objects - has proven to be a powerful tool (Newman, 2010). In recent years, substantial progress has been made by augmenting 'traditional' network theory in order to account for, e.g., the time-evolution of networks, the multiplex nature of many networks, networks of networks, and multiple types of connections between objects (Boccaletti et al., 2014; Kivelä et al., 2014; Berlingerio et al., 2011; Han, 2012; De Domenico et al., 2013; Gao et al., 2011a; Gao et al., 2011b; Santiago and Benito, 2008). However, even in its latest and most general form (i.e., multilayer networks (Kivelä et al., 2014)), network theory is still lacking crucial traits to serve as a general data analysis framework, and to bridge the gap between (big) data and its modelling. Most importantly, these include an explicit association of information with the nodes and edges of a network, and a conclusive representation of groups of nodes as well as the interactions between such groups on different scales. We consider these qualities to be indispensable, not only in the representation of complex systems, but also in their analysis. First, because they facilitate the means to represent features, relations and interactions on different scales, and second, because they allow us to coarse-grain, simplify and highlight important large-scale structures in a data-driven analysis.

To solve these issues is the primary contribution of this dissertation. We will provide a network-based framework that is capable of representing heterogeneous complex systems across scales. For the first time, groups of objects (supernodes) and their respective interrelations and interactions (superedges) are incorporated into a self-contained network representation. Furthermore, potentially unstructured and diverse information is explicitly associated with the different (super)nodes and (super)edges. For these reasons, our framework is capable of acting as a go-between, joining a unified and generalized network representation of systems with the tools and methods of traditional fields, such as multivariate statistics and probability theory, as well as the rising field of machine learning. Therefore, in combination with the software package that accompanies this dissertation, our *deep graphs* framework makes an important contribution to the field of complex systems and quite possibly to data analytics in general.

A number of applications will be presented, demonstrating benefits of the deep graph framework. In an explorative analysis of extreme rainfall measurements, we construct a deep graph to track and categorize the formation of spatio-temporal rainfall clusters. Thereby, we uncover propagation patterns over subtropical South America that were just recently discovered using rather complicated statistical methods, as well as extreme rainfall clusters over tropical South America that have not yet been identified and analyzed in the meteorological literature. Based on the constructed rainfall deep graph, we will also provide statistical evidence that the spatio-temporally integrated size distribution of extreme rainfall clusters does not - as previously suggested - follow a powerlaw. Instead, we find that the size distribution over the oceans is best approximated by an exponentially truncated powerlaw. Arguing with a generative storm-track model, we explain how the exponential truncation of the observed distribution could be caused by the presence of land masses. In another application of the deep graph framework, we combine two high-resolution satellite products in order to identify spatio-temporal clusters of fire-affected areas in the Brazilian Amazon and characterize their land-use specific burning conditions. By means of the statistical characteristics we find, we will take the first steps towards a probabilistic classifier of fire-clusters into land use types with the ultimate goal of predicting whether a measured fire-cluster was caused by anthropogenic activities or natural causes. Finally, we investigate the effects of white noise and global coupling strength on the maximum degree of synchronization for a variety of oscillator models coupled linearly and non-linearly according to a broad spectrum of network topologies. We find a general sigmoidal scaling and validate it with a simple regression model.

The dissertation is structured as follows. In chapter 2 we introduce the theoretical framework, *deep graphs*, laying the groundwork for this dissertation. In chapter 3 we employ the deep graph framework to track, cluster and categorize local formations of extreme rainfall. A statistical analysis of the size distribution of spatio-temporal extreme rainfall clusters is conducted in chapter 4. Chapter 5 is dedicated to a statistical characterization of fire-cluster burning conditions on different land use types in the Brazilian Amazon. In chapter 6 we investigate the effects of white noise and global coupling strength on the maximum degree of synchronization in complex networks. Finally, conclusions are drawn in chapter 7.

Part I.

Theoretical Framework

Chapter 2.

Deep Graphs - A General Framework to Represent and Analyze Heterogeneous Complex Systems across Scales

2.1. Summary

Network theory has proven to be a powerful tool in describing and analyzing systems by modelling the relations between their constituent objects. Particularly in recent years, great progress has been made by augmenting ‘traditional’ network theory in order to account for the multiplex nature of many networks, multiple types of connections between objects, the time-evolution of networks, networks of networks and other intricacies. However, existing network representations still lack crucial traits to serve as a general data analysis tool, and to bridge the gap between (big) data and its modelling. These include, most importantly, an explicit association of information with possibly heterogeneous types of objects and relations, and a conclusive representation of the properties of groups of nodes as well as the interactions between such groups on different scales. In this thesis, we introduce a collection of definitions resulting in a framework that, on the one hand, entails and unifies existing network representations (e.g., network of networks, multilayer networks), and on the other hand, generalizes and extends them by incorporating the above features. To implement these features, we first specify the nodes and edges of a finite graph as sets of properties (which are permitted to be arbitrary mathematical objects). Second, the mathematical concept of partition lattices is transferred to network theory in order to demonstrate how partitioning the node and edge set of a graph into supernodes and superedges allows to aggregate, compute and allocate information on and between arbitrary groups of nodes. The derived partition lattice of a graph, which we denote by *deep graph*, constitutes a concise, yet comprehensive representation that enables the expression and analysis of heterogeneous properties, relations and interactions on all scales of a complex system in a self-contained manner. Furthermore, to be able to utilize existing network-based methods and models, we derive the different representations of multilayer networks from our framework and demonstrate the advantages of our representation. We also provide a powerful software implementation of the theoretical framework introduced here, which integrates seamlessly into the PyData ecosystem making it accessible to a vast number of computational scientists. This chapter is

based on the associated publication P2, and the following sections will closely follow parts of this publication.

2.2. Introduction

We propose a framework that is capable of representing arbitrarily complex systems in a self-contained manner, and establishes an interface for the tools and methods developed in a variety of research disciplines, such as multivariate statistics, machine learning and graph theory. The framework is based on the ontological assumption that every system can be described in terms of its constituent objects (anything conceivable, i.e., "beings", "things", "entities", "events", "agents", "concepts" or "ideas") and their relations. With this assumption in mind, we build this framework based on graph or network theory. A graph, in its simplest form, is a collection of nodes (representing objects) where some pairs of nodes are connected by edges (representing the existence of a relation) (Bollobas, 1998). On top of that, we define an additional structure in order to meet the following objectives:

1. any node of the network may explicitly incorporate properties of the object(s) it represents. We refer to these properties as the *features* of a node, which themselves are mathematical objects.

2. any edge of the network may explicitly incorporate properties of the relation(s) it represents. We refer to these properties as the *relations* of an edge, which themselves are mathematical objects.

3. any subset of the set of all nodes of the network may be grouped into a *supernode*. Thereby, we may aggregate the features of the supernodes' constituent nodes. Furthermore, we may allocate features particular to that supernode ("emergent" properties of the compound supernode), based on either the aggregated features, a priori knowledge, or both.

4. any subset of edges of the set of all edges of the network may be grouped into a *superedge*. Thereby, we may aggregate the relations of the superedges' constituent edges. Furthermore, we may allocate relations particular to that superedge ("emergent" properties of the compound superedge), based on either the aggregated relations, a priori knowledge, or both.

5. we may place edges between any pair of supernodes, as well as between supernodes and nodes.

We believe that a comprehensive treatment of groups of objects, as well as their relations, is just as indispensable as an explicit incorporation of data, not only in the representation of complex systems, but also in their analysis. First, because it facilitates the means to represent features, relations and interactions on different scales, and second, because it allows us to coarse-grain, simplify and highlight important large-scale structures in a data-driven analysis.

Needless to say, this is not the first attempt to augment simple graphs in order to satisfy at least some of the above objectives. In weighted graphs, for instance, one can assign a number to each edge (i.e., the weight, strength, or distance of an edge) (Horvath, 2014). In node-weighted networks, it is possible to assign numbers to the nodes of the network (Wiedermann et al., 2013). In hypergraphs, one can define edges joining more than two vertices at a time (called hyperedges), essentially allowing for the assignment of groups in a network (Berge, 1976). Such a membership of nodes in groups can also be represented by bipartite networks, where one of two kinds of nodes represents the original objects, and the other kind represents the groups to which the objects belong (Asratian et al., 1998). Particularly in recent years – certainly also due to the deluge of available data – a multitude of frameworks has been proposed with the aim of pluralizing the number of labels and values that may be assigned to a node, and allowing for different categories of connections between pairs of nodes, such as, e.g.: multivariate networks; multidimensional networks; interacting networks; interdependent networks; networks of networks; heterogeneous information networks; and multilayer networks (see Boccaletti et al. (2014), Kivelä et al. (2014), Berlingerio et al. (2011), Han (2012), De Domenico et al. (2013), Gao et al. (2011a), Gao et al. (2011b), and Santiago and Benito (2008) and references therein).

However, none of these frameworks satisfies all of the above objectives at the same time. In contrast, the framework proposed in this thesis meets all these objectives. This allows us, on the one hand, to derive all of the above network representations as special cases by imposing certain constraints on our framework, which enables the utilization of the network-based methods, models and measures developed for them. On the other hand, we will demonstrate how the implementation of these objectives into our framework generalizes existing network representations, making it possible to combine heterogeneous datasets (e.g., climatological and socio-ecological data or (electro-)physiological records of different organs), integrate a priori knowledge of groups of objects and their relations, and conduct an analysis of potential interrelations of the respective systems within the same network representation. Based on the introduced framework, we also provide a Python software package that is fully scalable and integrates into the PyData ecosystem comprised of various libraries for scientific computing.

This chapter is structured as follows. The theoretical part of our framework is described in Sec. 2.3, where we introduce our representation of a graph, and Sec. 2.4, where we demonstrate a comprehensive manner of graph partitioning. Then, we outline the general procedure of constructing a deep graph in Sec. 2.5. We show how to impose traditional graph representations and how our framework integrates with existing data analysis tools in Secs. 2.6 and 2.7. Thereafter, we make a number of general remarks regarding the identification of nodes, edges, their respective properties and partitions (Sec. 2.8), and briefly describe the accompanying software package (Sec. 2.9). Finally, we draw our conclusions in Sec. 2.10.

Explanation	Symbol	Given by	Properties
# nodes	n	$\lvert V\rvert$	≥ 1
# supernodes	n^p	$\lvert V^p\rvert$	$1 \leq n^p \leq n$
# supernodes (IP)	$n^{\underline{p}}$	$\lvert V^{\underline{p}}\rvert$	$1 \leq n^{\underline{p}} \leq n$
# nodes in supernode i	$n^{p,i}$	$\lvert V_i^p\rvert$	$1 \leq n^{p,i} \leq n$
# nodes in supernode $\underline{i}$ (IP)	$n^{\underline{p},\underline{i}}$	$\lvert V_{\underline{i}}^{\underline{p}}\rvert$	$0 \leq n^{\underline{p},\underline{i}} \leq n$
# types of features in supernode $\underline{i}$ (IP)	$n_{\text{types}}^{\underline{p},\underline{i}}$	$\lvert T_{\underline{i}}^{\underline{p}}\rvert$	$0 \leq n_{\text{types}}^{\underline{p},\underline{i}} \leq \lvert F_{\underline{i}}^{\underline{p}}\rvert$
# features of type t in supernode $\underline{i}$ (IP)	$n_t^{\underline{p},\underline{i}}$	$\lvert F_{\underline{i},t}^{\underline{p},T}\rvert$	$\leq n^{\underline{p},\underline{i}}$
# partition-specific types of features in supernode $\underline{i}$ (IP)	${}^{\underline{p}}n_{\text{types}}^{\underline{p},\underline{i}}$	allocation	≥ 0
# features in node i	f_i	$\lvert V_i\rvert$	≥ 0
# distinct types of features in G	n_{types}	$\lvert T_v\rvert$	≥ 0
# features of type t in G	n_t	$\lvert F_{i,t}^{p^c,T}\rvert$	$\leq n$
# edges	m	$\lvert E\rvert$	≥ 0
# superedges	m^p	$\lvert E^p\rvert$	$0 \leq m^p \leq m$
# superedges (IP)	$m^{\underline{p}}$	$\lvert E^{\underline{p}}\rvert$	$0 \leq m^{\underline{p}} \leq m$
# edges in superedge (i, j)	$m^{p,ij}$	$\lvert E_{ij}^p\rvert$	$0 \leq m^{p,ij} \leq m$
# edges in superedge $(\underline{i}, \underline{j}, \underline{r})$ (IP)	$m^{\underline{p},\underline{ij},\underline{r}}$	$\lvert E_{\underline{ij},\underline{r}}^{\underline{p}}\rvert$	$0 \leq m^{\underline{p},\underline{ij},\underline{r}} \leq m$
# types of relations in superedge $(\underline{i}, \underline{j}, \underline{r})$ (IP)	$m_{\text{types}}^{\underline{p},\underline{ij},\underline{r}}$	$\lvert T_{\underline{ij},\underline{r}}^{\underline{p}}\rvert$	$0 \leq m_{\text{types}}^{\underline{p},\underline{ij},\underline{r}} \leq \lvert R_{\underline{ij},\underline{r}}^{\underline{p}}\rvert$
# relations of type t in supernode $(\underline{i}, \underline{j}, \underline{r})$ (IP)	$m_t^{\underline{p},\underline{ij},\underline{r}}$	$\lvert R_{\underline{ij},\underline{r},t}^{\underline{p},T}\rvert$	$\leq m^{\underline{p},\underline{ij},\underline{r}}$
# partition-specific types of relations in superedge $(\underline{i}, \underline{j}, \underline{r})$ (IP)	${}^{\underline{p}}m_{\text{types}}^{\underline{p},\underline{ij},\underline{r}}$	allocation	≥ 0
# relations in edge (i, j)	r_{ij}	$\lvert E_{ij}\rvert$	≥ 0
# distinct types of relations in G	m_{types}	$\lvert R_{ij}^{p^c}\rvert$	≥ 0
# relations of type t in G	m_t	$\lvert R_{ij,t}^{p^c,T}\rvert$	$\leq m$

Table 2.1.: Deep Graph Notation. The symbol "#" reads: "number of", and "IP" reads: "intersection partition".

2.3. Graph Representation

Throughout this chapter, we assume (w.l.o.g.) that (super)nodes, (super)edges, types of features and types of relations are represented by consecutive integers starting from 1. Also, in Tab. 2.1 one may find a summary of all the important quantities of a deep graph.

The basis of our representation is a finite, directed graph (possibly with self loops), given by a pair

$$G = (V, E), \tag{2.1}$$

where V is a set of $n := |V|$ nodes,

$$V = \{V_i \,|\, i \in \{1, 2, ..., n\}\}, \tag{2.2}$$

and E is a set of $m := |E|$ directed edges, given by

$$E \subseteq \{E_{ij} \,|\, i, j \in \{1, 2, ..., n\}\} =: E'. \tag{2.3}$$

Every node $V_i \in V$ of this graph represents some object(s), and every edge $E_{ij} \in E$ represents the existence of some relation(s) from node V_i to node V_j. We say that an edge E_{ij} is *incident* to both nodes V_i and V_j. In order to explicitly incorporate information or data of the objects and their pairwise relations, we specify every node V_i and every edge E_{ij} of G as a set of its respective properties. We refer to the properties of a node as its *features*, and to the properties of an edge as its *relations*.

Hence, we define every node V_i as a set of f_i features (and its index, to guarantee uniqueness of the nodes), given by

$$V_i = \{i, F_i^1, F_i^2, ..., F_i^{f_i}\}. \tag{2.4}$$

As opposed to the 'weight' of a node in node-weighted networks (Wiedermann et al., 2013) – which is usually a real number – a feature F_i^j can be any mathematical object (e.g. numbers; quantitative or categorical variables; sets; matrices; tensors; functions; nodes; edges; graphs; but also strings to represent abstract objects, such as concepts or ideas). Furthermore, we associate every feature with a *type*, in order to express the kind of property a feature is related to and to establish a comparability between the features of different nodes. For example, for a node representing a city, some types of features might be 'location', 'age', 'number of inhabitants', 'unemployment rate' and 'voting patterns'. For a node representing a neuron, some types of features could be 'time series of the membrane potential', 'measuring device' and 'distribution of ion channel types'. On that account, we denote with $F = \{F_i^j \,|\, i \in \{1, 2, ..., n\} \wedge j \in \{1, 2, ..., f_i\}\}$ the set of all features, and with $T_v = \{1, 2, ..., n_{\text{types}}\}$ the set of all distinct types of features contained in the graph G. We then define a surjective function mapping every feature to its corresponding type,

$$t_v : F \to T_v, F_i^j \mapsto t_v(F_i^j) := T_i^j \in T_v, \tag{2.5}$$

such that $t_v(F_i^j) = t_v(F_k^l)$ for all pairs of features that share the same type. However, we do not allow a node V_i to have multiple features of the same type, $t_v(F_i^j) \neq t_v(F_i^k)$ for all $j \neq k \in \{1, 2, ..., f_i\}$. In other words, every node V_i has exactly f_i distinct types of features. Figure 2.1(a) depicts different nodes along with their features and the feature's types.

Analogously, we define every edge $E_{ij} \in E$ as a set of r_{ij} relations (and its index pair, to guarantee uniqueness of the edges), given by

$$E_{ij} = \{(i, j), R_{ij}^1, R_{ij}^2, ..., R_{ij}^{r_{ij}}\}. \tag{2.6}$$

Again, as opposed to the commonly real-valued 'weight' of an edge in edge-weighted networks (Horvath, 2014), a relation R_{ij}^k can be any mathematical object. Just like features, we map every relation to its corresponding type, indicating the kind of property of a relation (e.g. 'distance between', 'correlation between', 'similarity between', 'works for', 'is part of'). We denote with $R = \{R_{ij}^k \,|\, i, j \in \{1, 2, ..., n\} \wedge k \in \{1, 2, ..., r_{ij}\}\}$ the set of all relations, and with $T_e = \{1, 2, ..., m_{\text{types}}\}$ the set of all distinct types of relations contained in the graph G. We then map every relation

a)

i	category	name	occupation	tax	age	income	gender	political ideology
0	person	Marge Gunderson	state police officer	n/a	46	50	female	egalitarianism
1	person	Frank Underwood	president	n/a	21	1200	male	conservatism
2	person	Avon Barksdale	chief executive officer	n/a	91	23000	male	anarchism
3	person	Leah Gould	judge	n/a	24	85	female	egalitarianism
4	person	Valeria Velez	chief editor	n/a	33	45	female	environmentalism
5	person	Mr. Garrison	teacher	n/a	29	35	male	environmentalism
6	person	Rust Cohle	federal police officer	n/a	61	40	male	egalitarianism
7	newspaper	Twin Peaks Gazette	n/a	n/a	198	n/a	n/a	conservatism
8	state	Krakozhia	n/a	state tax, 10%	198	n/a	n/a	conservatism
9	state	Tomania	n/a	state tax, 5%	141	n/a	n/a	conservatism
10	party	Ingsoc	n/a	n/a	198	n/a	n/a	conservatism
11	company	Viktor's Gun Shop	n/a	n/a	141	n/a	n/a	egalitarianism
12	think tank	Policy Solutions Association	n/a	n/a	198	n/a	n/a	conservatism
13	bank	Commonwealth Shared Risk	n/a	n/a	141	n/a	n/a	neoliberalism

b)

i	j	lives in	works for	monthly cash flow	kind of payment
0	8	TRUE	TRUE	5	tax
1	8	TRUE	TRUE	120	tax
1	9	n/a	TRUE	0	tax
2	0	n/a	n/a	60	bribe
2	1	n/a	n/a	500	bribe
2	2	n/a	TRUE	n/a	n/a
2	3	n/a	n/a	90	bribe
2	6	n/a	n/a	70	bribe
2	9	TRUE	n/a	20	tax
2	10	n/a	n/a	250	donation
2	11	n/a	n/a	20	expense
2	13	n/a	n/a	10000	investment
3	8	TRUE	TRUE	9	tax
3	9	n/a	TRUE	0	tax
4	7	n/a	TRUE	n/a	n/a
4	9	TRUE	n/a	3	tax
4	12	n/a	TRUE	5	donation
5	9	TRUE	TRUE	1	tax
5	10	n/a	TRUE	1	donation
6	8	TRUE	TRUE	4	tax
6	9	n/a	TRUE	0	tax
13	2	n/a	n/a	5000	investment

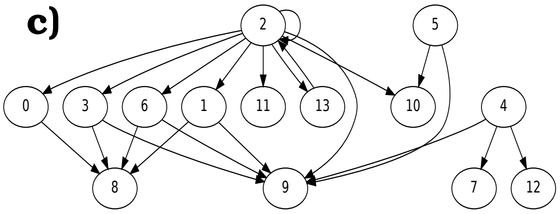


Figure 2.1.: Graph Representation. Illustration of a fictional graph $G = (V, E)$ consisting of $n = 14$ nodes and $m = 22$ directed edges. (a) Representation of the nodes $V_i \in V$. The left column indicates the nodes' indices, the top row indicates the *types* of features. A feature denoted "n/a" means that the corresponding node does not have a feature of the corresponding type. **(b)** Representation of the edges $E_{ij} \in E$. The first two columns from the left indicate the indices of edges E_{ij} from node V_i to node V_j and the top row indicates the *types* of relations. A relation denoted "n/a" means that the corresponding edge does not have a relation of the corresponding type. **(c)** Depiction of the graph's topology, where nodes are represented by indexed circles and edges are represented by arrows.

onto a type,

$$t_e : R \to T_e, R_{ij}^k \mapsto t_e(R_{ij}^k) := T_{ij}^k \in T_e, \tag{2.7}$$

such that $t_e(R_{ij}^k) = t_e(R_{mn}^l)$ for all pairs of relations that share the same type, and $t_e(R_{ij}^k) \neq t_e(R_{ij}^l)$ for all $k \neq l \in \{1, 2, ..., r_{ij}\}$. Therefore, every edge E_{ij} has exactly r_{ij} distinct types of relations. Figure 2.1(b) illustrates several edges with different types of relations.

For notational convenience later on, we define all elements $E_{ij} \in E'$ that are not in E as empty sets,

$$E_{ij} := \varnothing \text{ for all } i, j \in \{1, 2, ..., n\} : E_{ij} \in E' \setminus E. \tag{2.8}$$

In other words, we say an edge from node V_i to node V_j exists if $E_{ij} \neq \varnothing$, and in this context, we term V_i the *source* node and V_j the *target* node. Therefore, we can rewrite the set of edges of $G = (V, E)$ as

$$E = \{E_{ij} \,|\, i, j \in \{1, 2, ..., n\} \wedge E_{ij} \neq \varnothing\}. \tag{2.9}$$

2.4. Graph Partitioning

In this section, we introduce a comprehensive concept of graph partitioning. To avoid confusion: we do not refer to graph partitioning in the sense of finding "good" partitions (i.e. communities) based on some cost function or statistical measures (Buluç et al., 2013) such as, e.g., Newman's modularity measure (Newman and Girvan, 2004). Instead, we refer to graph partitioning in the more general sense of partitions of sets (Lucas, 1990).

First, we demonstrate how partitioning the node set V of a graph $G = (V, E)$ enables us to group arbitrary nodes into *supernodes*, and equivalently, how partitioning the edge set E allows us to group arbitrary edges into *superedges*. Then, we introduce a coherent manner of partitioning a graph $G = (V, E)$ into a *supergraph*, where the edge set E is partitioned in accordance with a given partition of the node set V, based on the edges' incidences to the nodes.

Partitioning nodes, edges or graphs – as we will show – not only conserves the information contained in the graph $G = (V, E)$, but allows us to redistribute it. This enables us to aggregate the features and relations of any desirable group of nodes and edges, and to allocate information particular to them. Furthermore, it facilitates the means to place edges between any supernodes or between supernodes and nodes, allowing us to represent interactions or relations on any scale of a complex system.

2.4.1. Partitioning Nodes

Given a graph $G = (V, E)$ with $n = |V|$ nodes, we define a surjective function mapping every node $V_i \in V$ to a supernode label (i.e. feature) ${}^vS_i$,

$$ {}^vp : V \to {}^vS = \{1, 2, ..., n^p\}, V_i \mapsto {}^vp(V_i) := {}^vS_i \in {}^vS. \tag{2.10}$$

This function induces a partition V^p of V into $n^p = |V^p|$ supernodes V_i^p, given by

$$V_i^p = \{V_j \,|\, j \in \{1, 2, ..., n\} \wedge {}^vp(V_j) = {}^vS_i\}, \text{ and} \tag{2.11}$$
$$V^p = \{V_i^p \,|\, i \in \{1, 2, ..., n^p\}\}. \tag{2.12}$$

The number of nodes a supernode $V_i^p \in V^p$ contains is denoted by $n^{p,i} := |V_i^p| \geq 1$.

The supernode labels given by the function ${}^vp(V_i) = {}^vS_i$ can be transferred as features to the nodes of G,

$$V_i = \{i, F_i^1, F_i^2, ..., F_i^{f_i}, {}^vS_i\}, \tag{2.13}$$

where the type of feature of ${}^vS_i$ is the same for all nodes, $t_v({}^vS_i) = t_v({}^vS_j)$ for all $i, j \in \{1, 2, ..., n\}$. In turn, every feature itself can be interpreted as a supernode label, and we can say that its corresponding type induces a partition of the node set. For instance, looking at Fig. 2.1(a), we see that the type of feature 'political ideology' induces a partition of V into $n^p = 5$ supernodes: 'egalitarianism' (consisting of $n^{p,1} = 4$ nodes), 'conservatism' ($n^{p,2} = 6$ nodes), 'anarchism' ($n^{p,3} = 1$ node), 'environmentalism' ($n^{p,4} = 2$ nodes) and 'neoliberalism' ($n^{p,5} = 1$ node). Since some nodes might not have a feature of a certain type [see for instance the type 'gender' in Fig. 2.1(a)], there is a degree of freedom when partitioning by that type: we can create one supernode comprising all nodes without the feature; create a separate supernode for every node without the feature; or create no supernode at all for these nodes. This choice is of course dependent on the analysis.

2.4.2. Partitioning Edges

Partitioning the edge set E of a given graph $G = (V, E)$ with $n = |V|$ nodes and $m = |E|$ edges can be realized just like partitioning the node set. However, since edges E_{ij} are incident to pairs of nodes (V_i, V_j), we later demonstrate how to exploit this association in order to partition edges based on properties of the nodes. Here, we demonstrate the procedure analogous to that of partitioning nodes. Hence, we define a surjective function mapping every edge $E_{ij} \in E$ to a superedge label (i.e. relation) ${}^eS_r$, given by

$$ {}^ep : E \to {}^eS = \{1, 2, ..., m^p\}, E_{ij} \mapsto {}^ep(E_{ij}) := {}^eS_r \in {}^eS. \tag{2.14}$$

This function induces a partition E^p of E into $m^p = |E^p|$ superedges E_r^p, where

$$E_r^p = \{E_{uv} \,|\, \Phi^e(u, v) \wedge {}^ep(E_{uv}) = {}^eS_r\}, \tag{2.15}$$

$$\Phi^e(u,v) : (u, v \in \{1,2,...,n\} \wedge E_{uv} \neq \varnothing), \text{ and} \tag{2.16}$$

$$E^p = \{E^p_r \,|\, r \in \{1,2,...,m^p\}\}. \tag{2.17}$$

The number of edges a superedge $E^p_r \in E^p$ contains is denoted by $m^{p,r} := |E^p_r| \geq 1$.

Equivalently to supernode labels, we can transfer the superedge labels given by the function ${}^e p(E_{ij}) = {}^e S_r$ as relations to the edges of G,

$$E_{ij} = \{(i,j), R^1_{ij}, R^2_{ij}, ..., R^{r_{ij}}_{ij}, {}^e S_r\}, \tag{2.18}$$

where the type of relation of ${}^e S_r$ is the same for all edges, $t_e({}^e S_i) = t_e({}^e S_j)$ for all $i, j \in \{1,2,...,m\}$. Again, every relation itself can be interpreted as a superedge label, and we say that its corresponding type induces a partition of the edge set. Looking at Fig. 2.1(b), we see that the type of relation 'kind of payment' induces a partition of E into $m^p = 6$ superedges: 'bribe' (consisting of $m^{p,1} = 4$ edges), 'donation' ($m^{p,2} = 3$ edges), 'expense' ($m^{p,3} = 1$ edge), 'investment' ($m^{p,4} = 2$ edges), 'tax' ($m^{p,5} = 10$ edges), and 'n/a' ($m^{p,6} = 2$ edges). Since the last two edges do not have a relation of the type 'kind of payment', we could have also partitioned the edges into $m^p = 5$ superedges (leaving the two edges out), or into $m^p = 7$ superedges (the two edges are put into separate superedges).

2.4.3. Partitioning a Graph

Here, we introduce a coherent manner of partitioning a graph $G = (V, E)$ with $n = |V|$ nodes and $m = |E|$ edges, based on the edges' incidences to the nodes. Given a partition V^p of V induced by a function ${}^v p(V_i) = {}^v S_i$ [see Eqs. (2.10)-(2.12)], we define the *corresponding* partition E^p of E into $m^p = |E^p|$ superedges E^p_{ij} by the following equations:

$$E^p_{ij} := \{E_{uv} \,\big|\, \Phi^e(u,v) \wedge {}^v p(V_u) = {}^v S_i \wedge {}^v p(V_v) = {}^v S_j\}, \tag{2.19}$$

where

$$\Phi^e(u,v) : (u, v \in \{1,2,...,n\} \wedge E_{uv} \neq \varnothing)\,, \text{ and} \tag{2.20}$$

$$E^p := \{E^p_{ij} \,\big|\, i, j \in \{1,2,...,n^p\} \wedge E^p_{ij} \neq \varnothing\}. \tag{2.21}$$

By this definition, we group all edges E_{ij} originating from nodes in supernode V^p_i and targeting nodes in supernode V^p_j into a superedge E^p_{ij}, consisting of $m^{p,ij} := |E^p_{ij}| \geq 0$ edges. It is straightforward to show that this corresponding partition is indeed a partition of E, and therefore we can say that partitioning the node set V by ${}^v p$ induces a *supergraph* $G^p = (V^p, E^p)$. In reference to the graph in Fig. 2.1, a partition of the nodes by the type of feature 'category', for instance, would yield a supergraph consisting of $n^p = 7$ supernodes: 'bank' (consisting of $n^{p,1} = 1$ node), 'company' ($n^{p,2} = 1$ node), 'newspaper' ($n^{p,3} = 1$ node), 'party' ($n^{p,4} = 1$ node), 'person' ($n^{p,5} = 7$ nodes), 'state' ($n^{p,6} = 2$ nodes) and 'think tank' ($n^{p,7} = 1$ node); and $m^p = 8$ corresponding superedges: from 'bank' to 'person' (consisting of $m^{p,15} = 1$

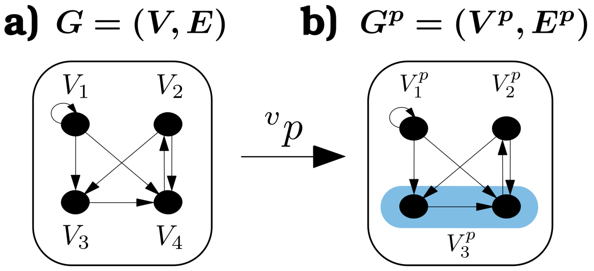


Figure 2.2.: Graph Partitioning. Illustration of a supergraph, 'naturally' induced by a partition of the node set. (a) The graph $G = (V, E)$, comprised of $n = 4$ nodes $V = \{V_1, V_2, V_3, V_4\}$ and $m = 7$ edges $E = \{E_{11}, E_{13}, E_{14}, E_{23}, E_{24}, E_{34}, E_{42}\}$. **(b)** The supergraph $G^p = (V^p, E^p)$, obtained by grouping the nodes V_3 and V_4 into the supernode $V_3^p = \{V_3, V_4\}$. It is comprised of $n^p = 3$ nodes $V^p = \{V_1^p, V_2^p, V_3^p\}$, and $m^p = 5$ edges E^p, given by: $E_{11}^p = \{E_{11}\}, E_{13}^p = \{E_{13}, E_{14}\}, E_{23}^p = \{E_{23}, E_{24}\}, E_{32}^p = \{E_{42}\}$, and $E_{33}^p = \{E_{34}\}$.

edge); and from 'person' to: 'bank' ($m^{p,51} = 1$ edge), 'company' ($m^{p,52} = 1$ edge), 'newspaper' ($m^{p,53} = 1$ edge), 'party' ($m^{p,54} = 2$ edges), 'person' ($m^{p,55} = 5$ edges), 'state' ($m^{p,56} = 10$ edges) and 'think tank' ($m^{p,57} = 1$ edge). See also Fig. 2.2 for an illustration of grouping a graph's nodes and edges into a supergraph.

2.4.4. The Partition Lattices of a Graph

In this section, we explain some general mathematical properties that arise when partitioning a graph. This provides for a deeper understanding of this framework, and sets the stage for the next sections.

Before we go into details of graph-specific partitioning, we point out some relevant properties of what in mathematics is known as *partition lattices* (Birkhoff, 1940). Assume we are given a finite, non-empty n-element set X. The total number of distinct partitions we can create of it is given by the Bell number $B(n)$ (Bell, 1938; Becker and Riordan, 1948). The set of all possible partitions, which we denote by $P = \{P_i \,|\, i \in \{1, 2, ..., B(n)\}\}$, is a *partially ordered* set, since some of the elements of P have a pair-wise relation, which is called the *finer-than* relation. A partition P_i is said to be a *refinement* of a partition P_j, if every element of P_i is a subset of some element of P_j. If this condition is fulfilled, one says that P_i is *finer* than P_j, $P_i \leq P_j$, and vice versa, P_j is *coarser* than P_i, $P_j \geq P_i$. Since X is finite, every partition P_i is bounded from below and from above with respect to this finer-than relation,

$$P_f \leq P_i \leq P_c, \text{ for all } i \in \{1, 2, ..., B(n)\}, \tag{2.22}$$

where P_f is called the *finest* element of P, given by $P_f = \{\{X_1\}, \{X_2\}, ..., \{X_n\}\}$, and P_c is the *coarsest* element, given by the trivial partition $P_c = \{X\}$. This implies that each set of elements of P has a finest upper bound and a coarsest lower bound. Therefore, the set of all possible partitions P is called a partition lattice (or more precisely, a *geometric* lattice, since X is finite (Welsh, 2010)). Any *totally* ordered subset of P is called a *chain*, and any subset of P for which there exists no relation between any two different elements of that subset is called an *antichain*.

Since in this thesis we are dealing with finite graphs exclusively, we can directly build the lattices of the node set V and the edge set E, and translate the above properties of lattices into the context of graphs. However, we will also make use of the natural way of partitioning a graph as demonstrated in Sec. 2.4.3, in order to create the geometric lattice of a graph $G = (V, E)$. This lattice, by construction, entails the lattice of the node set V, and a specific subset of the lattice of the edge set E, and there are therefore two lattices of interest: the lattice of a graph G, and the lattice of its edges E.

Let us note down the lattice of a graph G with n nodes and m edges, for which there is a total of $B(n)$ different supergraphs. We create the set of all distinct partitions of V by prescribing a set of functions ${}^{v}p = \{{}^{v}p^k \,|\, k \in \{1, 2, ..., B(n)\}\}$, such that each function

$$ {}^{v}p^k : V \to {}^{v}S^k = \{1, 2, ..., n^{p^k}\}, \tag{2.23} $$

$$ V_i \mapsto {}^{v}p^k(V_i) := {}^{v}S_i^k \in {}^{v}S^k, k \in \{1, 2, ..., B(n)\}, \tag{2.24} $$

induces a supergraph $G^{p^k} = (V^{p^k}, E^{p^k})$ as demonstrated in Sec. 2.4.3 and illustrated in Fig. 2.2. The partition lattice of V, induced by the set of functions ${}^{v}p$, is therefore given by ${}^{V}L = \{V^{p^k} \,\big|\, k \in \{1, 2, ..., B(n)\}\}$. The finer-than relation between partitions translated to the lattice of V means that if $V^{p^k} \leq V^{p^l}$, then every supernode $V_i^{p^l}$ of V^{p^l} is the union of supernodes $V_j^{p^k} \in V^{p^k}$. We transfer the finer-than relation to graphs, by saying that $G^{p^k} \leq G^{p^l}$, if both $V^{p^k} \leq V^{p^l}$ and $E^{p^k} \leq E^{p^l}$. With reference to Eqs. (2.19)-(2.21), we see that for all partitions $V^{p^k} \leq V^{p^l}$, it follows that $E^{p^k} \leq E^{p^l}$ by construction, and consequently, we denote with ${}^{G}L = \{G^{p^k} \,\big|\, k \in \{1, 2, ..., B(n)\}$ the partition lattice of G, henceforth referred to as the *deep graph* of G. The lattice of the graph depicted in Fig. 2.2(a) is illustrated in Fig. 2.3. Some of its properties are: the finest element of ${}^{G}L$ is the graph $G = (V, E)$ itself; the coarsest element, which we denote by $G^{p^c} = (V^{p^c}, E^{p^c})$, consists of one supernode connected to itself by a single superedge; and every chain in ${}^{G}L$, illustrated by the red, dashed lines in Fig. 2.3, corresponds to some agglomerative, hierarchical clustering of the nodes of G. The dashed blue lines in Fig. 2.3 will be explained in the next section.

However, the lattice of E is generally not covered entirely by the lattice of G. In fact, maximally $B(n)$ of $B(m)$ possible partitions of E are contained in ${}^{G}L$, due to the partitioning of E by correspondence [see Eqs. (2.19)-(2.21)]. The full lattice of E can be created analogously to that of V, by prescribing a set of functions ${}^{e}p = \{{}^{e}p^k \,|\, k \in \{1, 2, ..., B(m)\}\}$, such that each function

$$ {}^{e}p^k : E \to {}^{e}S^k = \{1, 2, ..., m^{p^k}\}, \tag{2.25} $$

$$ E_{ij} \mapsto {}^{e}p^k(E_{ij}) := {}^{e}S_r^k \in {}^{e}S^k, k \in \{1, 2, ..., B(m)\}, \tag{2.26} $$

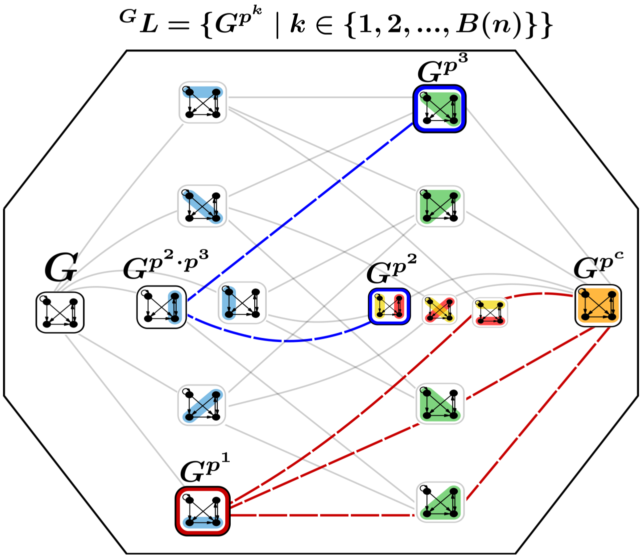


Figure 2.3.: A deep graph, i.e. the geometric partition lattice of a graph. Illustration of the graph $G = (V, E)$ as described in Fig. 2.2(a), and its $B(n) = 15$ corresponding supergraphs, ordered by refinement from the right to the left. The supergraph G^{p^1} is illustrated in detail in Fig. 2.2(b). Each link in this Hasse diagram corresponds to the finer-than relation between a pair of supergraphs. The dashed lines colored in red correspond to chains in the lattice. An intersection partition is illustrated by $G^{p^2 \cdot p^3}$, which results from intersecting G^{p^2} and G^{p^3}. It constitutes a refinement of both G^{p^2} and G^{p^3} (blue dashed lines). The figure is a modification of Wikimedia Commons (2015).

induces a partition E^{p^k} of E as demonstrated in Eqs. (2.14)-(2.17). The lattice of E is then given by ${}^{E}L = \{E^{p^k} \,|\, k \in \{1, 2, ..., B(m)\}\}$.

In the next section, we introduce a useful tool that can be used to navigate the lattices ${}^{G}L$ and ${}^{E}L$ for the sake of creating meaningful partitions, based on the features and relations of a given graph.

2.4.5. Intersection Partitions

Due to the rapid increase of possible partitions with growing numbers of nodes and edges, it is only possible to actually compute the full lattices of G and E for very small graphs. However, we are generally not interested in every single partition, but rather a meaningful subset of them. Here, we demonstrate how to create *intersection partitions* and thereby establish a valuable tool to find potentially informative partitions based on the features and relations of a graph. Furthermore, as we will demonstrate later, one can utilize intersection partitions in order to compute similarity measures between different partitions. We will also make use of intersection partitions in Sec. 2.6 in order to derive a tensor-like representation of a multilayer network (De Domenico et al., 2013).

To begin with, let us demonstrate what we mean by intersection partitions with a simple example. Imagine a standard 52-card deck, partitioned by color on the one hand (red and black, both comprised of 26 cards), and by suit on the other hand (spades, diamonds, hearts and clubs, each comprised of 13 cards). The intersection partition of color and suit would then be comprised of 8 elements: cards that are red and at the same time spades (0 cards); red & diamonds (13 cards); etc. Before showing some examples with regard to the exemplary graph in Fig. 2.1, let us note down the different ways of creating intersection partitions of a graph.

We first demonstrate the construction of intersection partitions of V. Assume we are given a set of K $[\leq B(n)]$ partitions of V, induced by a set of functions ${}^{v}p = \{{}^{v}p^k \,|\, k \in I^K\}$, where $I^K = \{1, 2, ..., K\}$ is the partition index set. From this set of available partitions, we choose a collection $g \subseteq I^K$, which is used to create an intersection partition. We define an element $V^{\underline{p}}_{\underline{i}}$ of the intersection partition $V^{\underline{p}}$ by

$$V^{\underline{p}}_{\underline{i}} := \{V_j \,\big|\, j \in \{1, 2, ..., n\} \wedge \forall k \in g : {}^{v}p^k(V_j) = {}^{v}S^k_{i^k}\}, \text{ where} \tag{2.27}$$

$$\underline{p} = (p^k)_{k \in g}, \underline{i} = (i^k)_{k \in g}, i^k \in \{1, 2, ..., n^{p^k}\}, \tag{2.28}$$

and the intersection partition itself by

$$V^{\underline{p}} := \bigcup_{\underline{i}} V^{\underline{p}}_{\underline{i}}. \tag{2.29}$$

Since $\varnothing \notin V^{\underline{p}}$ by construction, and by showing that

$$V^{\underline{p}}_{\underline{i}} \cap V^{\underline{p}}_{\underline{j}} = \varnothing \text{ for all } V^{\underline{p}}_{\underline{i}} \neq V^{\underline{p}}_{\underline{j}} \in V^{\underline{p}}, \text{ where} \tag{2.30}$$

$$\underline{j} = (j^k)_{k \in g}, j^k \in \{1, 2, ..., n^{p^k}\}, \tag{2.31}$$

we see that $V^{\underline{p}}$ is indeed a partition of V. A supernode $V^{\underline{p}}_{\underline{i}}$ of an intersection partition is comprised of $n^{\underline{p},\underline{i}} := |V^{\underline{p}}_{\underline{i}}|$ nodes V_j of G that simultaneously belong to all supernodes ${}^vS^k_{i^k}$ chosen by g. The number of supernodes of an intersection partition, $n^{\underline{p}} := |V^{\underline{p}}|$, is bounded by $\prod_{k \in g} n^{p^k}$, and every intersection partition constitutes a refinement of the partitions it has been constructed from, $V^{\underline{p}} \leq V^{p^k}$ for all $k \in g$. The number of distinct intersection partitions we can construct from I^K is bounded from above by $I(K) = \sum_{|g|=0}^{K} \binom{K}{|g|} = |\mathcal{P}(I^K)|$, where $\mathcal{P}(I^K)$ is the power set of the partition index set, hence $I(K) = 2^K$. Looking at Fig. 2.1, the intersection partition of the collection of partitions $g = \{\text{‘category’}, \text{‘political ideology’}\}$ would yield $n^{\underline{p}} = 10$ supernodes. The supernodes comprised of more than 1 node of G would be: ‘person’ & ‘egalitarianism’ (3 nodes); ‘person’ & ‘environmentalism’ (2 nodes); and ‘state’ & ‘conservatism’ (2 nodes).

Defining the corresponding intersection partition $E^{\underline{p}}$ of E into $m^{\underline{p}} := |E^{\underline{p}}|$ superedges $E^{\underline{p}}_{\underline{ij}}$ by

$$E^{\underline{p}}_{\underline{ij}} := \{E_{uv} \,\Big|\, \Phi^e(u,v) \wedge \forall k \in g : {}^vp^k(V_u) = {}^vS^k_{i^k} \wedge \forall k \in g : {}^vp^k(V_v) = {}^vS^k_{j^k}\} \tag{2.32}$$

and $E^{\underline{p}} := \bigcup_{\underline{i},\underline{j}} E^{\underline{p}}_{\underline{ij}}$, it follows that $\underline{p}$ induces a supergraph $G^{\underline{p}} = (V^{\underline{p}}, E^{\underline{p}})$, in analogy to Eqs. (2.19)-(2.21). A superedge $E^{\underline{p}}_{\underline{ij}}$ is comprised of $m^{\underline{p},\underline{ij}} := |E^{\underline{p}}_{\underline{ij}}|$ edges E_{uv} originating from nodes in supernode $V^{\underline{p}}_{\underline{i}}$ and targeting nodes in supernode $V^{\underline{p}}_{\underline{j}}$. Figure 2.3 depicts a supergraph, created from intersecting two different supergraphs (blue dashed lines).

With regard to partitioning the edges of a graph, however, there are other options than partitioning by types of relations [see Eqs. (2.14)-(2.17)], or by correspondence [see Eqs. (2.19)-(2.21)]. We now show how to utilize the edges' relations and the features of their incident nodes in all possible combinations. For instance, regarding the graph in Fig. 2.1, we might want to know how many edges originate from nodes with a ‘political ideology’ of ‘egalitarianism’, or ‘conservatism’, etc. The answer would yield a total of $m^p = 5$ superedges, originating from: ‘anarchism’ (comprised of 9 edges); ‘egalitarianism’ (5 edges); ‘conservatism’ (2 edges); ‘environmentalism’ (5 edges); and ‘neoliberalism’ (1 edge). These superedges, however, could be refined by asking how many of their constituent edges target nodes of the ‘category’ ‘bank’, or ‘company’ and so forth. We would then see, for instance, that the edges originating from nodes with a ‘political ideology’ of ‘egalitarianism’ all target nodes of the ‘category’ ‘state’. Refining these superedges even further, we could ask, how many edges originating from nodes with a ‘political ideology’ of ‘egalitarianism’ and targeting nodes of the ‘category’ ‘state’ are of the ‘kind of payment’ ‘tax’, or ‘bribe’, etc. Let us note down all the combinations formally, to clarify the procedure of partitioning edges.

Assume we are given a set of ${}^vK$ partitions of V, induced by ${}^vp = \{{}^vp^k \,|\, k \in {}^vI^K\}$, where ${}^vI^K = \{1, 2, ..., {}^vK\}$ is the partition index set of the nodes. Additionally, we have a set of ${}^eK$ partitions of E, induced by ${}^ep = \{{}^ep^k \mid k \in {}^eI^K\}$, where

${}^{e}I^{K} = \{1, 2, ..., {}^{e}K\}$ is the partition index set of the edges. From these partitions, we choose three different collections: a *source type collection* $g^{s} \subseteq {}^{v}I^{K}$, a *target type collection* $g^{t} \subseteq {}^{v}I^{K}$ and a *relation type collection* $g^{r} \subseteq {}^{e}I^{K}$. Then, we denote a superedge by $E^{\underline{p}}_{\underline{i}\underline{j},\underline{r}}$, where

$$\underline{p} = \left(({}^{s}p^{k})_{k \in g^{s}}, ({}^{t}p^{k})_{k \in g^{t}}, ({}^{r}p^{k})_{k \in g^{r}}) \right), \tag{2.33}$$

$$\underline{i} = (i^{k})_{k \in g^{s}}, \text{ with } i^{k} \in \{1, 2, ..., n^{p^{k}}\}, \tag{2.34}$$

$$\underline{j} = (j^{k})_{k \in g^{t}}, \text{ with } j^{k} \in \{1, 2, ..., n^{p^{k}}\}, \text{ and} \tag{2.35}$$

$$\underline{r} = (r^{k})_{k \in g^{r}}, \text{ with } r^{k} \in \{1, 2, ..., m^{p^{k}}\}, \tag{2.36}$$

and define it by

$$E^{\underline{p}}_{\underline{i}\underline{j},\underline{r}} := \{E_{uv} \,\Big|\, \Phi^{e}(u,v) \wedge \Phi^{v}_{g^{s}}(u) \wedge \Phi^{v}_{g^{t}}(v) \wedge \Phi^{e}_{g^{r}}(u,v)\}, \text{ where} \tag{2.37}$$

$$\Phi^{v}_{g^{s}}(u) : (\forall k \in g^{s} : {}^{v}p^{k}(V_{u}) = {}^{v}S^{k}_{i^{k}}), \tag{2.38}$$

$$\Phi^{v}_{g^{t}}(v) : (\forall k \in g^{t} : {}^{v}p^{k}(V_{v}) = {}^{v}S^{k}_{j^{k}}), \text{ and} \tag{2.39}$$

$$\Phi^{e}_{g^{r}}(u,v) : (\forall k \in g^{r} : {}^{e}p^{k}(E_{uv}) = {}^{e}S^{k}_{r^{k}}). \tag{2.40}$$

The partition $E^{\underline{p}}$ of E is then given by $E^{\underline{p}} := \bigcup_{\underline{i},\underline{j},\underline{r}} E^{\underline{p}}_{\underline{i}\underline{j},\underline{r}}$. Based on these definitions, we denote the number of superedges by $m^{\underline{p}} = |E^{\underline{p}}|$, and the number of edges contained in a superedge by $m^{\underline{p},\underline{i}\underline{j},\underline{r}} := |E^{\underline{p}}_{\underline{i}\underline{j},\underline{r}}|$. If all collections are empty at the same time, $g^{x} = \varnothing$ for all $x \in \{s, t, r\}$, it follows that $E^{\underline{p}}_{\underline{i}\underline{j},\underline{r}} = E$, which means that the edge set is partitioned into the trivial partition, comprised of one superedge entailing all edges. Furthermore, if we choose $g^{s} = g^{t}$ and $g^{r} = \varnothing$, we get the definition of the corresponding partition, as stated in Eq. (2.32). Expressed formally, the example stated in the above paragraph would hence be described as follows: we choose the source type collection by $g^{s} = \{$'political ideology'$\}$, the target type collection by $g^{t} = \{$'category'$\}$ and the relation type collection by $g^{r} = \{$'kind of payment'$\}$. The superedge $E^{\underline{p}}_{\underline{i}\underline{j},\underline{r}}$ corresponding to $\underline{i} =$ ('egalitarianism'), $\underline{j} =$ ('state') and $\underline{r} =$ ('tax') would then be comprised of $m^{\underline{p},\underline{i}\underline{j},\underline{r}} = 5$ edges.

Before we turn to the next section, let us make some general remarks regarding intersection partitions:

i) First of all, it is noteworthy that it only makes sense to create intersection partitions of antichains, since any chain in g, g^{s}, g^{t} or g^{r} can be replaced by the finest element of the respective chain.

ii) When creating intersection partitions we have to be aware of the fact that a supernode $V^{\underline{p}}_{\underline{i}}$ might be comprised of zero nodes, $n^{\underline{p},\underline{i}} = 0$. In this case, we say the supernode $V^{\underline{p}}_{\underline{i}}$ does not exist. This stands in contrast to the supernodes $V^{p^{k}}_{i}$ of supergraphs $G^{p^{k}}$, for which $n^{p,i} \geq 1$ for all $i \in \{1, 2, ..., n^{p^{k}}\}$, since we chose the functions ${}^{v}p^{k}$ to be surjective. This does not pose a problem though, since for a superedge $E^{\underline{p}}_{\underline{ij}}$ with $m^{\underline{p},\underline{ij}} = 0$, we can still deduce if the superedge does not exist because at least one of the supernodes does not exist ($n^{\underline{p},\underline{i}}$ or $n^{\underline{p},\underline{j}} = 0$), or because there is in fact no superedge between existing supernodes ($n^{\underline{p},\underline{i}}$ and $n^{\underline{p},\underline{j}} \geq 1$).

iii) Finally, we want to refer to Appendix A, where we demonstrate how to utilize intersection partitions in order to compute similarity measures between different (intersection) partitions. Such measures can be utilized, for instance, to assess the community structure of time-evolving networks, as Granell et al. (2015) have demonstrated.

2.4.6. Redistribution and Allocation of Information on the Lattices

The last sections were dedicated to constructing partitions, allowing us to group any desirable subset of nodes and edges into supernodes and superedges, respectively. Here, we demonstrate that the information of a graph – expressed by the features and relations of its constituent nodes and edges – is not only conserved under partitioning, but redistributed on the partition lattices, according to the partition function(s) we choose. This allows us to aggregate data of any desirable group of nodes or edges. We then demonstrate how to allocate partition-specific features and relations, which also allows us to create superedges independently of the edges in G.

First, the information contained in a given graph $G = (V, E)$ is conserved when creating partitions: given a partition $V^{\underline{p}}$ of V induced by $\underline{p}$ [see Eqs. (2.27)-(2.29)], every supernode $V^{\underline{p}}_{\underline{i}} \in V^{\underline{p}}$ is a subset of the nodes of G, where each node is comprised of a set of features. The complete set of features contained in supernode $V^{\underline{p}}_{\underline{i}}$ can then be partitioned by their corresponding types, and therefore expressed as a collection of sets of features of common type. Hence, a supernode $V^{\underline{p}}_{\underline{i}}$ – expressed in terms of its constituent features – is given by

$$V^{\underline{p}}_{\underline{i}} = \{\underline{i}\} \cup \{F^{\underline{p},T}_{\underline{i},t}\}_{t \in \{1,2,\dots,n^{\underline{p},\underline{i}}_{\text{types}}\}}, \tag{2.41}$$

where the number of distinct types of features in supernode $V^{\underline{p}}_{\underline{i}}$ is denoted by $n^{\underline{p},\underline{i}}_{\text{types}}$, and the number of features of type t by $n^{\underline{p},\underline{i}}_{\mathrm{t}} := |F^{\underline{p},T}_{\underline{i},t}|$. Looking at Fig. 2.1, the supernode comprised of the nodes with indices $(2, 6, 11)$, for instance, has a total of $n^{\underline{p},\underline{i}}_{\text{types}} = 7$ types of features: ‘category’ ($n^{\underline{p},\underline{i}}_{1} = 3$: ‘person’: 2 nodes, ‘company’: 1 node); ‘name’ ($n^{\underline{p},\underline{i}}_{2} = 3$: ‘Avon Barksdale’: 1 node, ‘Rust Cohle’: 1 node, ‘Viktor’s Gun Shop’: 1 node); ‘occupation’ ($n^{\underline{p},\underline{i}}_{3} = 2$: ‘chief executive officer’: 1 node, ‘federal police officer’: 1 node); ‘age’ ($n^{\underline{p},\underline{i}}_{4} = 3$: ‘91’: 1 node, ‘61’: 1 node, ‘141’: 1 node); etc. By this example, it becomes clear that we can easily create frequency distributions of the values of a supernodes’ different types of features.

Analogously, we can express superedges in terms of the relations of their constituent edges, which we also partition by their corresponding types: given a partition $E^{\underline{p}}$ of

E induced by $\underline{p}$ [see Eqs. (2.33)-(2.40)], a superedge $E^{\underline{p}}_{\underline{ij},\underline{r}}$ is given by

$$E^{\underline{p}}_{\underline{ij},\underline{r}} = \{(\underline{i},\underline{j},\underline{r})\} \cup \{R^{\underline{p},T}_{\underline{ij},\underline{r},t}\}_{t\in\{1,2,\dots,m^{\underline{p},\underline{ij},\underline{r}}_{\text{types}}\}}, \tag{2.42}$$

where the number of distinct types of relations in superedge $E^{\underline{p}}_{\underline{ij},\underline{r}}$ is denoted by $m^{\underline{p},\underline{ij},\underline{r}}_{\text{types}}$, and the number of relations of type t by $m^{\underline{p},\underline{ij},\underline{r}}_{\text{t}} := |R^{\underline{p},T}_{\underline{ij},\underline{r},t}|$. For mathematical details, we refer to Appendix B.

By this representation of supernodes and superedges, we can clearly see that the information of a graph G is not only conserved under partitioning, but redistributed according to the partition function(s) we choose. This means that every supergraph $G^{\underline{p}}$ on the lattice ${}^{G}L$, and every partition $E^{\underline{p}}$ on the lattice ${}^{E}L$, corresponds to a unique redistribution of the information contained in a graph G, and the collection of all possible redistributions is given by the lattices ${}^{G}L$ and ${}^{E}L$.

Second, we show how to allocate partition-specific information on the lattice ${}^{G}L$. Note that we omit the vector notation of intersection partitions for the remainder of this section for reasons of notational simplicity. Given a supergraph $G^p \in {}^{G}L$, we know that its supernodes are comprised of features $\{F^{p,T}_{i,t}\}_{t\in\{1,2,\dots,n^{p,i}_{\text{types}}\}}$, and its superedges are comprised of relations $\{R^{p,T}_{ij,t}\}_{t\in\{1,2,\dots,m^{p,ij}_{\text{types}}\}}$. Based on these features and relations, we can compute additional properties (e.g., moments, correlations) by applying some set of functions on them. For the sake of notational convenience, we write single functions mapping to sets of new properties:

$$f(\{F^{p,T}_{i,t}\}_{t\in\{1,2,\dots,n^{p,i}_{\text{types}}\}}) = \{{}^{p}F^{p,T}_{i,t}\}_{t\in\{1,2,\dots,{}^{p}n^{p,i}_{\text{types}}\}}, \tag{2.43}$$

$$f(\{R^{p,T}_{ij,t}\}_{t\in\{1,2,\dots,m^{p,ij}_{\text{types}}\}}) = \{{}^{p}R^{p,T}_{ij,t}\}_{t\in\{1,2,\dots,{}^{p}m^{p,ij}_{\text{types}}\}}, \tag{2.44}$$

where the additional p-index on the upper left corner indicates that these features and relations are specific to the supergraph G^p. Of course, we can also allocate features to supernodes independently from the features of the supernodes' constituent nodes. The same goes for the relations of superedges, even in the case when they are comprised of zero edges (for which $E^p_{ij} = \varnothing$ and therefore also $\{R^{p,T}_{ij,t}\}_{t\in\{1,2,\dots,m^{p,ij}_{\text{types}}\}} = \varnothing$). We do not, however, denote these independently allocated features and relations differently to the computed features and relations in Eqs. (2.43) and (2.44). Hence, the properties of supernodes and superedges of a supergraph can be written as

$$V^p_i = \{i\} \cup \{F^{p,T}_{i,t}\}_{t\in\{1,2,\dots,n^{p,i}_{\text{types}}\}} \cup \{{}^{p}F^{p,T}_{i,t}\}_{t\in\{1,2,\dots,{}^{p}n^{p,i}_{\text{types}}\}}, \tag{2.45}$$

and

$$E^p_{ij} = \{(i,j)\} \cup \{R^{p,T}_{ij,t}\}_{t\in\{1,2,\dots,m^{p,ij}_{\text{types}}\}} \cup \{{}^{p}R^{p,T}_{ij,t}\}_{t\in\{1,2,\dots,{}^{p}m^{p,ij}_{\text{types}}\}}. \tag{2.46}$$

Of course, the partition-specific features and relations only bear meaning for the unique element of the lattice $G^p \in {}^GL$. Furthermore, they can only be redistributed on the set of all coarser supergraphs, given by the chains entailed in $\{G^{p'} \mid p' \in \{1, 2, ..., B(n)\} \wedge p' > p\} \subseteq {}^GL$ (see the red, dashed lines in Fig. 2.3).

2.5. How to Construct a Deep Graph

The theoretical framework satisfying the objectives stated in the Introduction is now fully described. Here, we want to roughly describe the general procedure of constructing a deep graph. For this purpose, we introduce two types of auxiliary functions: *connectors*, which are functions allowing us to create (super)edges between (super)nodes, purely based on the properties of the represented objects; and *selectors*, which are functions allowing us to select (i.e. filter) (super)edges, based on their respective properties. In combination, these functions effectively allow us to forge the topology of a deep graph, which we will exemplify in chapters 3 and 5. Furthermore, we demonstrate in this section how our framework integrates with existing network theory and other data analysis tools, and finally make some general remarks regarding the identification of (super)nodes, (super)edges and partitions.

Given a set of n objects, the general procedure of constructing a deep graph can be outlined as follows

1. identify each object as a node V_i, $i = 1, 2, ..., n$.
2. assign features to each node V_i, $V_i = \{i, F_i^1, F_i^1, ..., F_i^{f_i}\}$.
3. define *connectors*

   $$m_{ij}(V_i, V_j) := E_{ij} = \{(i, j), R_{ij}^1, R_{ij}^2, ..., R_{ij}^{r_{ij}}\}, \tag{2.47}$$

   where m_{ij} is a function mapping a pair of sets of features to a set of relations. Connector functions create "computable", or "external" relations between objects. They are typically based on distance or similarity measures of objects, or some information or physical flow between them. A few examples are the scalar product of vectors, the distance of objects in a metric space, or correlation coefficients between variables. Networks solely based on one such measure are often termed functional networks (Boers et al., 2013; Zhou et al., 2006).
4. create the set of all possible edges E' by applying the connector functions on all pairs of nodes.
5. if there is any a priori knowledge of relations between the objects (as opposed to the computed relations by connectors), append them to the corresponding edges. By a priori known relations, we mean any inherent, internal, physical, trivial or abstract relations, such as flightpaths between airports, synapses between neurons, social relationships between humans, or relations of plants to the treatment of medical conditions.

6. define *selectors*

$$s_{ij}(E_{ij}) := \begin{cases} E_{ij} & \text{if } E_{ij} \text{ satisfies conditions of } s_{ij} \\ \varnothing & \text{if } E_{ij} \text{ does not satisfy conditions of } s_{ij} \end{cases}, \tag{2.48}$$

   where s_{ij} is a function mapping a set of relations to itself, if the set satisfies the conditions expressed in the function, or to an empty set otherwise, thereby removing the corresponding edge from the edge set E. Selectors can be simple thresholding functions (e.g., for some features F_j^k and F_i^l: $E_{ij} \mapsto E_{ij}$ if $|F_j^k - F_i^l| \leq T$, else $E_{ij} \mapsto \varnothing$), but they can also be more complicated and elaborate, involving different types of relations at the same time.

7. select $E \subseteq E'$ by applying the selector functions on all edges E'.

The graph is then given by $G = (V, E)$, where the objects' properties are represented by sets of features V_i, and the relational information of pairs of objects is represented by sets of relations E_{ij}.

The next step is to repeat the following procedure for any supergraph $G^p \in {}^G L$ for which we want to allocate, aggregate or evaluate information. Again, for notational clarity, we omit vector notation.

1. identify a partition G^p of G. This partition might be induced by the (intersection of) features of the nodes in G (see Sec. 2.4.1 and Sec. 2.4.5), or created by any other means, such as manual assignment of supernode labels, clustering algorithms, community detection algorithms, or partitioning by the connected components of G.

2. compute and allocate partition-specific features to any of the supernodes

$$V_i^p = \{i\} \cup \{F_{i,t}^{p,T}\}_{t \in \{1,2,\dots,n_{\text{types}}^{p,i}\}} \cup \{{}^p F_{i,t}^{p,T}\}_{t \in \{1,2,\dots,{}^p n_{\text{types}}^{p,i}\}}.$$

3. compute and allocate partition-specific relations to any of the superedges

$$E_{ij}^p = \{(i,j)\} \cup \{R_{ij,t}^{p,T}\}_{t \in \{1,2,\dots,m_{\text{types}}^{p,ij}\}} \cup \{{}^p R_{ij,t}^{p,T}\}_{t \in \{1,2,\dots,{}^p m_{\text{types}}^{p,ij}\}}.$$

4. define *connectors* between supernodes,

$$m_{ij} : V^p \times V^p \to E'^p, (V_i^p, V_j^p) \mapsto m_{ij}(V_i^p, V_j^p),$$

   to further enrich the relations of the superedges in G^p,

$$E_{ij}^p = \{(i,j)\} \cup \{R_{ij,t}^{p,T}\}_{t \in \{1,2,\dots,m_{\text{types}}^{p,ij}\}} \cup \{{}^p R_{ij,t}^{p,T}\}_{t \in \{1,2,\dots,{}^p m_{\text{types}}^{p,ij}\}} \cup m_{ij}(V_i^p, V_j^p).$$

5. define *selectors* on the set of superedges,

$$s_{ij}(E^p_{ij}) := \begin{cases} E^p_{ij} & \text{if } E^p_{ij} \text{ satisfies conditions of } s_{ij} \\ \varnothing & \text{if } E^p_{ij} \text{ does not satisfy conditions of } s_{ij} \end{cases} .$$

6. select $E^p \subseteq E'^p$ by applying the selector functions on all edges E'^p.

The supergraph is then represented by $G^p = (V^p, E^p)$. Repeating this procedure for different elements $G^p \in {}^G L$, we continuously extend the information contained in ${}^G L$. This information, in turn, can then be redistributed on the lattice (see Sec. 2.4.6, and the red lines in Fig. 2.3), and increases the number of possible ways to create intersection partitions (see Sec. 2.4.5, and the blue lines in Fig. 2.3).

2.6. Imposing Traditional Graph Representations

Here, we show how to obtain existing network representations, by imposing certain restrictions on our framework resulting in the multilayer network (MLN) representation, as defined by Kivelä et al. (2014). We chose to demonstrate only the attainment of the MLN representation for two reasons. First, because it is – to the best of our knowledge – the most general framework of network representation today, and second, because it allows us refer to the extensive work done by Kivelä et al. (2014), Boccaletti et al. (2014), and references therein. In these papers, the reader can find derivations of many additionally constrained network representations down to the level of ordinary graphs (Bollobas, 1998), as well as a compendium of network tools, models and concepts to analyze networks. Therefore, the derivation of the MLN representation – in conjunction with the work done in these papers – allows us to exploit the already existing tool set of network theory.

For readers unfamiliar with MLNs, we provide a summary in Appendix A. Without loss of generality, we assume a MLN $M = (V_M, E_M, V^N, \boldsymbol{L})$ with $|V^N| =: N$ nodes and $|V_M| =: n \leq |V^N| \cdot \prod_{a=1}^{d} |L_a|$ node-layers. First, we have to restrict ourselves to the representation of a single element of the partition lattice of a deep graph, $G^p \in {}^G L$. Let us assume that this element is the finest element of ${}^G L$ w.l.o.g., $G = (V, E)$. Then, there are two choices of G, resulting in distinct representations of M. We can place the additional information attributed to the layered structure of M either in the nodes of G, or in the edges of G. The latter case is described in Appendix D. The former case, which is the favourable representation of M, is described in the following.

We identify each node $V_i \in V = \{V_1, V_2, ..., V_n\}$ with a node-layer $V_{M,i} \in V_M$, such that

$$V_i = \{V_i^N, L_{1,i}, L_{2,i}, ..., L_{d,i}\} \mathrel{\widehat{=}} V_{M,i} \in V^N \times L_1 \times L_2 \times \cdots \times L_d, \tag{2.49}$$

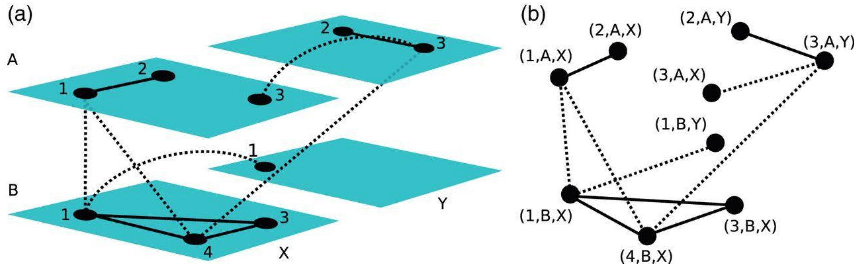
where $V_i^N \in V^N$ and $L_{a,i} \in L_a$ for all $a \in \{1, 2, ..., d\}$. This means that every node V_i of G has one feature corresponding to the index of a node $V_i^N \in V^N$ and

d features corresponding to elementary layers of the aspects $L_a \in \boldsymbol{L}$. An edge $E_{ij} \in E' = \{E_{11}, E_{12}, ..., E_{nn}\}$ is given by

$$E_{ij} = \begin{cases} \{w\left((V_{M,i}, V_{M,j})\right)\} & \text{if } (V_{M,i}, V_{M,j}) \in E_M \\ \varnothing & \text{if } (V_{M,i}, V_{M,j}) \notin E_M \end{cases}. \tag{2.50}$$

Therefore, the edge set E corresponding to E_M is given by $E = \{E_{ij} \mid i, j \in \{1, 2, ..., n\} \wedge E_{ij} \neq \varnothing\}$. Every edge $E_{ij} \in E$ has exactly one relation, whose type is determined by the tuple of features $(\{L_{a,i}\}_{a=1}^{d}, \{L_{a,j}\}_{a=1}^{d})$ of the adjacent nodes V_i and V_j. The derived representation $G = (V, E)$ corresponds one to one to the 'supra-graph' representation of a MLN, given by the tuple (V_M, E_M). Figure 2.4 shows an examplary MLN, side by side with its representation derived here and a tensor-like representation we derive in Appendix D.

In Appendix D, we demonstrate how the subset of the partition lattice ${}^{G}L$ of $G \mathrel{\widehat{=}} M$ induced by the types of features of its constituent nodes corresponds to different representations of a MLN, including the above mentioned tensor-like representation (De Domenico et al., 2013). There, we also discuss the constraints imposed on our framework in order to obtain the MLN representation, and how our representation solves the issues encountered in the representation of MLNs.



(c) $\boldsymbol{V}$

i	V^N	L_1	L_2
0	1	A	X
1	1	B	X
2	1	B	Y
3	2	A	X
4	2	A	Y
5	3	A	X
6	3	A	Y
7	3	B	X
8	4	B	X

(d) $\boldsymbol{E}$

i	j	R_i
0	1	r_0
0	3	r_1
0	8	r_2
1	2	r_3
1	7	r_4
1	8	r_5
4	6	r_6
5	6	r_7
6	8	r_8
7	8	r_9

(e) $\boldsymbol{E^{V^N \cdot L_1 \cdot L_2}}$

i^{V^N}	i^{L_1}	i^{L_2}	j^{V^N}	j^{L_1}	j^{L_2}	R_i
1	A	X	1	B	X	r_0
1	A	X	2	A	X	r_1
1	A	X	4	B	X	r_2
1	B	X	1	B	Y	r_3
1	B	X	3	B	X	r_4
1	B	X	4	B	X	r_5
2	A	Y	3	A	Y	r_6
3	A	X	3	A	Y	r_7
3	A	Y	4	B	X	r_8
3	B	X	4	B	X	r_9

Figure 2.4.: An exemplary multilayer network (MLN) and its representation by our framework. **(a)** An exemplary MLN, $M = (V_M, E_M, V^N, \boldsymbol{L})$, consisting of four nodes, $V^N = \{1, 2, 3, 4\}$, and two aspects, $\boldsymbol{L} = \{L_1, L_2\}$, where $L_1 = \{A, B\}$ and $L_2 = \{X, Y\}$. It has a total of 9 node-layers, $V_M = \{(1, A, X), (1, B, X), (1, B, Y), (2, A, X), (2, A, Y), (3, A, X), (3, A, Y), (3, B, X), (4, B, X)\}$, connected pair-wise by a total of 10 edges, $E_M \subset V_M \times V_M$. For notational brevity, we consider the edges to be directed (with randomly chosen directions). **(b)** The same MLN as in (a), depicted by its underlying 'supra-graph' representation, $G_M = (V_M, E_M)$. **(c)** The nodes $V_i \in V$ of the graph $G = (V, E)$, representing the MLN described in (a). G has a total of 9 nodes (corresponding to the MLN's node-layers), whose indices are given by the left column. The top row indicates the nodes' *types* of features, which correspond one-to-one to the MLN's node indices and its aspects. **(d)** The edges $E_{ij} \in E$ of the graph $G = (V, E)$, representing the MLN described in (a). G has a total of 10 edges, corresponding to the edges E_M of M. The first two columns indicate the indices of edges E_{ij} from node V_i to node V_j. The (complex- or real-valued) relations of the edges are denoted by r_i and their corresponding *types* by R_i (which are condensed into one column, for reasons of space). **(e)** A tensor-like representation of the edges of the MLN described in (a). It is derived from the graph G [see (c) and (d)], by constructing the intersection partition of all its *types* of features, resulting in the supergraph $G^{V^N \cdot L_1 \cdot L_2} = (V^{V^N \cdot L_1 \cdot L_2}, E^{V^N \cdot L_1 \cdot L_2})$. The supergraph's edges $E^{V^N \cdot L_1 \cdot L_2}_{i^{V^N} \cdot i^{L_1} \cdot i^{L_2}, j^{V^N} \cdot j^{L_1} \cdot j^{L_2}} \in E^{V^N \cdot L_1 \cdot L_2}$ are indexed like a tensor, as apparent from the table. See Appendix D for mathematical details. Figure 2.4(a) and (b) are reproduced with permission from Journal of Complex Networks 2, 203 - 271 (2014). Copyright 2013 Oxford University Press - Journals.

2.7. Integration with other Data Analysis Tools

As demonstrated above, the (super)nodes and (super)edges of a (super)graph $G^{\underline{p}} \in {}^{G}L$ are nothing less than collections of sets of mathematical objects,

$$V_{\underline{i}}^{\underline{p}} = \{\underline{i}\} \cup \{F_{\underline{i},t}^{\underline{p},T}\}_{t\in\{1,2,\ldots,n_{\text{types}}^{\underline{p},\underline{i}}\}} \cup \{{}^{\underline{p}}F_{\underline{i},t}^{\underline{p},T}\}_{t\in\{1,2,\ldots,{}^{\underline{p}}n_{\text{types}}^{\underline{p},\underline{i}}\}}, \text{ and} \tag{2.51}$$

$$E_{\underline{ij}}^{\underline{p}} = \{(\underline{i},\underline{j})\} \cup \{R_{\underline{ij},t}^{\underline{p},T}\}_{t\in\{1,2,\ldots,m_{\text{types}}^{\underline{p},\underline{ij}}\}} \cup \{{}^{\underline{p}}R_{\underline{ij},t}^{\underline{p},T}\}_{t\in\{1,2,\ldots,{}^{\underline{p}}m_{\text{types}}^{\underline{p},\underline{ij}}\}} \cup m_{\underline{ij}}(V_{\underline{i}}^{\underline{p}}, V_{\underline{j}}^{\underline{p}}), \tag{2.52}$$

just like the superedges of the partitions $E^{\underline{p}} \in {}^{E}L$ [see Eq. (2.42)]

$$E_{\underline{ij},\underline{r}}^{\underline{p}} = \{(\underline{i},\underline{j},\underline{r})\} \cup \{R_{\underline{ij},\underline{r},t}^{\underline{p},T}\}_{t\in\{1,2,\ldots,m_{\text{types}}^{\underline{p},\underline{ij},\underline{r}}\}}. \tag{2.53}$$

Therefore, there is nothing hindering us from utilizing the tool sets developed in fields such as multivariate statistics, probability theory, supervised and unsupervised machine learning, and graph theory, in order to analyze the properties of (super)nodes and their relations. For instance, we can use machine learning algorithms to predict missing features of nodes, or to predict relations between objects. We can use statistical tools to compute properties such as moments, ranges, covariances and cross-entropies. We can also compute graph theoretical measures, such as centrality measures (e.g. eigenvector centralities, betweenness centralities, closeness centralities, degree centralities), participation coefficients, matching indices or local clustering coefficients. Furthermore, we can compute similarity or distance measures through connector functions and then use appropriate clustering algorithms, such as stochastic block models, in order to find informative partitions (Aicher et al., 2013; Peixoto, 2012; Peixoto, 2013; Peixoto, 2014). All these properties and labels can then be reassigned to the features and relations of the (super)nodes and (super)edges.

2.8. Identification of (Super)Nodes, (Super)Edges and Partitions

The framework we have laid down offers a good deal of flexibility in mapping systems onto networks. For that reason, we want to conclude this section by making a number of general remarks regarding the identification of (super)nodes, (super)edges, their respective properties and partitions.

i) First of all, recall that the nodes of a graph represent arbitrary objects. There are no restrictions of what constitutes an object, so a node might represent literally anything that comes to mind. On top of that, the features of a node themselves can be arbitrary objects. This means, however, that the features of a node might themselves be identified as nodes, and vice versa. With regard to the exemplary graph in Fig. 2.1, for instance, the nodes with indices 8 and 9 (each representing a 'state') might just as well have been identified as features (of type 'lives in') of the nodes representing persons (indices 0-6). Yet, we identified them as nodes connected by edges (with the type of relation 'lives in') to the nodes 0-6. There are, of course, no general rules of what to identify as features, and what as nodes. This choice depends mainly on the context.

ii) A similar situation arises when dealing with variables $X = \{x_i \,|\, i \in \{1, 2, ..., n\}\}$. Imagine, for instance, a set of variables, each representing a time-series of measurements (e.g. of the channels in an EEG measurement). Then, each variable can be identified as a node, whose feature is the variable itself. But we could also identify each single value assumed by the variables as a node, and include features indicating the variables (supernodes) each node belongs to. Similar to the identification of the node-layers (as opposed to the nodes) of a MLN as the nodes of a deep graph (see Appendix B), the latter choice is more flexible, and actually contains the former choice as supernodes. By identifying each value of a time-series as a node, for instance, we can create additional supernode labels corresponding to discretizations of either axis (time or values), such as a discretization into a certain number of quantiles. Such a concept has been used by Campanharo et al. (2011) to create a map from a time series to a network with an approximate inverse operation. Within our framework, a bijection between a variable X and the nodes of a graph G is trivially given by

$$m_b : X \leftrightarrow V, x_i \mapsto m_b(x_i) := V_i = \{i, x_i\}, \tag{2.54}$$

$$m_b^{-1}(m_b[X]) = X. \tag{2.55}$$

Similarly, we can map multidimensional objects (or observations, in machine learning parlance)

$$X = \{\underline{x}_i = (x_i^j)_{j\in\{1,2,...,p\}} \in \mathbb{R}^p \,\big|\, i \in \{1, 2, ..., n\}\} \tag{2.56}$$

to the nodes of a graph $G = (V, E)$, by a function

$$m_b : X \leftrightarrow V, \underline{x}_i \mapsto m_b(\underline{x}_i) := V_i = \{i, \underline{x}_i\}, \tag{2.57}$$

$$m_b^{-1}(m_b[X]) = X. \tag{2.58}$$

This allows us, for instance, to create edges between objects containing the derivatives of each pair of variables, $m(V_i, V_j) := E_{ij} = \{\frac{x_j^k - x_i^k}{x_j^l - x_i^l}\}_{k \neq l \in \{1,2,...,p\}}$.

iii) It is also straightforward to represent and analyze recurrence networks (Marwan, 2008) by deep graphs. Given a p-dimensional phase space and a (discretized) phase-space trajectory represented by a temporal sequence of p-dimensional vectors $\underline{x}_i \in X$ [see Eq. 2.56], we first map each point $\underline{x}_i$ of the trajectory to a node V_i as described in Eq. 2.57. Then, we create edges between these nodes, based on some metric on the given phase space (e.g., the euclidian distance), $m(V_i, V_j) := E_{ij} = \{\|\underline{x}_j - \underline{x}_i\|\}$. Finally, we define a *selector* $s(E_{ij})$, $s(E_{ij}) \mapsto E_{ij}$ if $\|\underline{x}_j - \underline{x}_i\| < \varepsilon$, else $s(E_{ij}) \mapsto \varnothing$, leaving only edges between nodes with a distance smaller than ε, indicating the recurrence of a state in phase space. The recurrence network is then given by $G = (V, E)$. This approach can be generalized to cross and joint recurrence networks (Marwan and Kurths, 2002), by mapping a collection of phase space trajectories to the nodes of a graph (where to each node an additional feature is prescribed, indicating the trajectory it belongs to), and defining *connectors* and *selectors* accordingly.

iv) As a general rule of thumb, any divisible or separable entity of a system to be mapped to a deep graph should be divided into separate nodes, and their membership to the corresponding entity indicated by supernode labels.

v) A convenient manner of representing the time evolution of a network, for instance, is to take a graph (such as the one illustrated in Fig. 2.1), and prescribe to every node (edge) a feature (relation) of the type 'time'. Then, one simply copies the nodes and edges of the graph, indicates their point in time, and adjusts their features and relations according to whatever properties of the graph have changed over time. The deep graph incorporating the time evolution of the network is then given by joining all the copies of nodes and edges that we created into one graph.

vi) In terms of detecting partitions and identifying supernodes, we can also exploit the topological structure of a graph. In this respect, the auxiliary *connector* and *selector* functions introduced above constitute a helpful tool. Given a set of objects, the application of connectors and selectors allows us to effectively forge the topology of a (super)graph according to the research question at hand. This is particularly useful for spatially, temporally or spatio-temporally embedded systems, where we can define a metric space in which we place the objects of interest. Thereby, for instance, we may track objects in space over time by connecting them whenever they are close according to the metric, and then identify the connected components as the trajectories of the objects. Or, as we will demonstrate in the next chapter, we can use graph forging as a clustering scheme inducing a partition of the objects, and then define similarity measures on the induced subgraphs to detect recurrences of patterns.

vii) Finally, we want to emphasize that the identification of supernodes and superedges also constitutes a convenient manner of querying a deep graph, by allowing us to select any desirable group of nodes and edges, in order to aggregate their respective properties. Such a query could also involve graph theoretic objects, such as: in- and out-neighbours of a (super)node; paths; trees; forests; clusters; components; or communities.

2.9. The DeepGraph Software Package

The introduced framework is accompanied by a data analysis software package (The *DeepGraph Python Package*) [https://github.com/deepgraph/deepgraph]. The basis of this software package is Pandas, a fast and flexible data analysis tool for the Python programming language, and part of the *PyData Ecosystem* comprised of various libraries for scientific computing. Utilizing one of its primary data structures, the DataFrame (i.e., a table), we represent the (super)nodes of a graph by one set of tables, and their pairwise relations (i.e. the (super)edges of a graph) by another set of tables. Its main features are

- **create edges:** A method that enables an iterative, yet vectorized computation of pairwise relations (edges) between nodes using arbitrary, user-defined functions on the nodes' properties. The method provides arguments to parallelize the computation and control memory consumption, making it suitable for very large datasets and adjustable to whatever hardware you have at hand (from netbooks to cluster architectures).

- **partition nodes, edges or a graph:** Methods to partition nodes, edges or a graph by the graph's properties and labels, enabling the aggregation, computation and allocation of information on and between arbitrary *groups* of nodes. These methods also let you express elaborate queries on the information contained in a deep graph.

- **interfaces to other packages:** Methods to convert to common network representations and graph objects of popular Python network packages (e.g., SciPy sparse matrices, NetworkX graphs, graph-tool graphs).

- **plotting:** A number of useful plotting methods for networks, including drawings on geographical map projections.

2.10. Conclusions

We have introduced a collection of definitions resulting in *deep graphs*, a theoretical framework to describe and analyze heterogeneous systems across scales. Our framework unifies existing network representations and generalizes them by fulfilling two essential objectives: an explicit incorporation of information or data, and a comprehensive treatment of groups of objects and their relations. The former objective is implemented by specifying the nodes and edges of a (super)graph as sets of their respective properties. These properties, which may differ from node to node and from edge to edge, can be arbitrary mathematical objects. The second objective is implemented by transferring the mathematical concept of partition lattices to our graph representation. We have demonstrated how partitioning the node and edge set of a graph facilitates the means to aggregate, compute and allocate information on and between arbitrary groups of nodes. This information can then be stored on the lattices of a graph, allowing us to express and study properties, relations and interactions on all scales of the represented system(s).

Based on our representation, we were able to show how deep graphs establish an interface for common data analysis and modelling tools. This includes network-based concepts, models and methods, since we derived the different representations of a multilayer network (Kivelä et al., 2014), which was the most general network representation to date.

Yet, we have also introduced additional tools to support a comprehensive data analysis. We have demonstrated how the auxiliary *connector* and *selector* functions enable us to create and select (super)edges, thereby allowing us to forge the topology of a deep graph. Intersection partitions not only allow us to derive a tensor-like representation of a multilayer network (De Domenico et al., 2013), but also to calculate similarity measures between (intersection) partitions of a graph and to express elaborate queries on the information contained in a deep graph.

The *DeepGraph Python Package* (see Sec. 2.9) provides a powerful software implementation of the theoretical framework introduced here, and integrates seamlessly into the *PyData Ecosystem* making it accessible to a vast number of computational scientists.

We hope that our framework initiates attempts to generalize existing network measures and to develop new measures, particularly in respect of the heterogeneity of a system's components and their interactions on different scales. In the context of multilayer networks, generalizations of network measures have already led to significant new insights, and we expect the same to become true for deep graphs.

Part II.

Applications

Chapter 3.

Spatio-Temporal Tracking and Clustering of Extreme Rainfall

3.1. Summary

We exemplify an application of deep graphs using a dataset comprising 16 years of satellite-derived, gauge-calibrated global rainfall measurements. We extract extreme rainfall events from the data and construct a deep graph representation in order to track and categorize the formation of spatio-temporal rainfall clusters. First, we represent the extreme rainfall events as nodes of a graph, whose features indicate their location, time and rainfall rate. Creating edges on the basis of spatio-temporal proximity between nodes allows us to identify cohesive clusters as the connected components of the graph. Thereby, we can track and visualize the clusters' temporal evolution and calculate characterizing features, such as their lifetime, their spatial coverage, and the total volume of water they precipitated. We further agglomerate clusters into regional families based on a metric of spatial overlap between them. Finally, we discuss climatological characteristics of two of these families over the South American continent. The first, which is concentrated over the subtropics, was just recently discovered using rather complicated statistical methods, while the second, which is concentrated over tropical South America, has to our knowledge not yet been identified and analyzed in the meteorological literature. These particular clusters could be a promising subject for further research. The approach of tracking and coarse-graining spatio-temporal events applied here is easily transferable to other systems, as for instance demonstrated in chapter 5 of this thesis. This chapter is based on the associated publication P2, and some of the following sections will closely follow parts of this publication.

3.2. Introduction

In this chapter, we present a use case of the deep graph framework particularly relevant in the context of spatio-temporal systems. Given a set of events embedded in space and time, a common task is to cluster them into groups based on their spatio-temporal proximity. A number of algorithms exist for that purpose, such as: K-Means clustering (Arthur and Vassilvitskii, 2007), Mean-Shift clustering (Comaniciu and Meer, 2002), Spectral clustering (Ng et al., 2001) or DBSCAN (density-based spatial clustering of applications with noise) (Ester et al., 1996). Here, we introduce a novel, network-based clustering procedure relying on user-defined pairwise distance metrics and a set of thresholds. First, the events are represented as nodes of a graph, whose features indicate their spatio-temporal coordinates. Then, the pairwise proximity of events is determined by applying appropriate connector and selector functions (see Sec. 2.5). The connector functions express the distance metrics to be used, whereas the selector functions eliminate edges between events we deem not to be related by imposing thresholds on the spatial and temporal distances. Thereby, a topological structure is created which we use to identify clusters as the connected components of the network. An advantage of this clustering scheme is that the number of clusters does not have to be determined beforehand. However, suitable thresholds have to be selected in the context of the problem at hand, which indirectly determines the number of clusters.

Here, we use this clustering procedure to track the formation and propagation of extreme rainfall events in space and time. The basis is a quasi-global, high-resolution satellite product comprised of rainfall measurements from 1998 to 2014. Although other network-based studies of this dataset exist, they focus on the creation of synchronization-based functional networks from the time-series of rainfall measurements corresponding to the different geographical locations (e.g., Stolbova et al., 2014; Boers et al., 2013; Boers et al., 2014).

First, we identify spatio-temporal rainfall clusters and track their temporal resolution. Then, we partition the resulting clusters into regional families according to their spatial overlap. Finally, climatological interpretations of two exemplary propagation patterns over the South American continent are provided.

3.3. The Rainfall Data

We employ the Tropical Rainfall Measuring Mission (TRMM) 3B42 V7 dataset (Huffman et al., 2007), with 3-hourly temporal resolution for the time period from 1998 to 2014. The dataset is spatially gridded at a resolution of $0.25° \times 0.25°$ ranging from 50°S to 50°N. Each of the $N = 46.752 \cdot 1440 \cdot 400 \approx 2.69 \cdot 10^{10}$ data point consists of the time of the measurement t_i, the geographical location given by a tuple of coordinates (lon_i, lat_i), and the average rainfall rate during a 3-hour time window r_i.

We extract extreme rainfall events from the data by considering only those measurements above the 90th percentile of so-called *wet times* (defined as data points with rainfall rates $r \geq 0.1$ mm/h). The 90th percentile is chosen in agreement with the definition of extreme rainfall events in the IPCC report (Field et al., 2012a). The threshold values at each geographical location are depicted in Fig. 3.1. This results in $n \approx 2.16 \times 10^8$ extreme events, which serve as the data basis for the following construction of a deep graph.

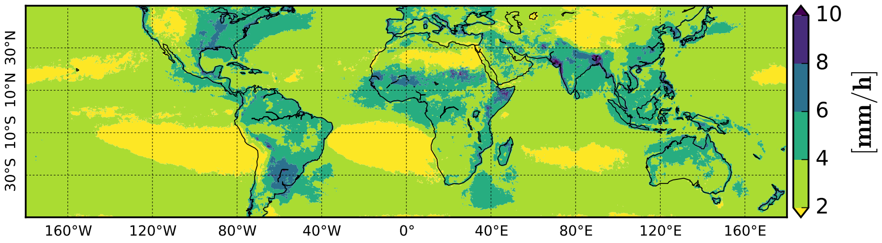


Figure 3.1.: Extreme rainfall thresholds. The $90th$ percentile threshold values (in [mm/h]) for each geographical location. Only rainfall events with rainfall rates above these thresholds are considered in this study.

3.4. Preprocessing and Representation of the Data as Nodes of a Graph

We first identify each of the n extreme rainfall events as a node V_i of the Graph $G = (V, E)$, with $V = \{V_i \,|\, i \in \{1, 2, ..., n\}\}$. Then, we assign features F_i^j to the nodes V_i by processing the information given by the dataset as follows.

We enumerate the given longitude, latitude and time coordinates, in order to associate every node with discrete space-time coordinates, $(lon_i, lat_i, t_i) \leftrightarrow (x_i, y_i, t_i) =: \underline{x}_i$. By this association, we are embedding the nodes into a 3-dimensional grid-cell geometry, which we will use below to identify spatio-temporal clusters. Furthermore, to each tuple (lon_i, lat_i) we assign a *geographical label*, $(lon_i, lat_i) \leftrightarrow L_i$, such that nodes with the same geographical location share the same label. This will enable us to measure spatial overlaps of spatio-temporal clusters later on. We also compute the surface area a_i and the volume of water precipitated v_i for each node. Hence, at this stage, every node has a total of six features, $V_i = \{L_i, \underline{x}_i, a_i, r_i, v_i\}$, as summarized in Tab. 3.1(a).

$$\boldsymbol{G = (V, E)}$$

a) $\boldsymbol{V_i}$

feature	symbol	type of feature	given by
F_i^1	L_i	geographical label	$(x_i, y_i) \leftrightarrow L_i$
F_i^2	$\underline{x}_i$	space-time coordinates	$(lon_i, lat_i, t_i) \leftrightarrow (x_i, y_i, t_i)$
F_i^3	a_i	surface area	$(111\text{km})^2 \cdot (0.25)^2 \cdot \cos\left(\frac{2\pi}{360°} \cdot lat_i\right)$
F_i^4	r_i	precipitation rate	given
F_i^5	v_i	vol. of water precipitated	$a_i \cdot r_i \cdot 3\text{h}$
F_i^6	C_i	cluster membership	connceted components
F_i^7	F_i	family membership	linkage clustering

b) $\boldsymbol{E_{ij}}$

relations	condition	symbol	type of relation	given by
R_{ij}^1	if $\lvert d\alpha_{ij}\rvert \leq 1 \; \forall\alpha \in \{x, y, t\}$	$d\underline{x}_{ij}$	spatio-temporal distance	$\underline{x}_j - \underline{x}_i$
$\emptyset$	else			

Table 3.1.: The features and relations of the graph $G = (V, E)$**. (a)** The features of the nodes V_i, representing extreme rainfall measurements. The type of feature 'cluster membership' is introduced in Sec. 3.5 and the type of feature 'family membership' in Sec. 3.6. **(b)** The relation of the edges E_{ij}, representing the spatio-temporal distance between rainfall measurements. An edge only exists, if the condition stated in the table is fulfilled.

3.5. Partitioning into Spatio-Temporal Rainfall Clusters

As we have assigned the same types of features to all nodes, we can define a single connector that we apply to all pairs of nodes,

$$m(V_i, V_j) := E_{ij} = \{(\underline{x}_j - \underline{x}_i)\} =: \{d\underline{x}_{ij}\}. \tag{3.1}$$

The set of all edges is therefore given by $E' = \{E_{ij} \,|\, i, j \in \{1, 2, ..., n\}\}$, where each of the $|E'| \approx 4.69 \cdot 10^{16}$ elements corresponds to a discrete distance vector of a pair of measurements. The edges of G will be utilized to detect spatio-temporal clusters in the data, or in more technical terms: to partition the set of all nodes into subsets of connected grid points. One can imagine the nodes to be elements of a 3 dimensional grid box, where we allow every node to have 26 possible neighbours (8 neighbours in the time slice of the measurement, t_i, and 9 neighbours in each the time slice $t_i - 1$ and $t_i + 1$). We can compute the clusters by identifying them as the connected components of the graph $G = (V, E)$, where E is given by applying the selector

$$s(E_{ij}) := \begin{cases} E_{ij} & \text{if } |d\alpha_{ij}| \leq 1 \forall\alpha \in \{x, y, t\} \wedge i \neq j \\ \varnothing & \text{else} \end{cases} \tag{3.2}$$

on all edges, such that $E = \{E_{ij} \,|\, i, j \in \{1, 2, ..., n\} \wedge E_{ij} \neq \varnothing\}$ leaves only $m = |E| \approx 9.16 \cdot 10^8$ edges between nodes that are neighbours on the grid. See Fig. 3.2 for an illustration of this clustering scheme.

Identifying the connected components of G results in a labelling of the nodes according to their respective cluster membership. We find a total of $n^C \approx 1.42 \cdot 10^7$ spatio-temporal clusters, and transfer their labels as features to the nodes of G, $V_i = \{L_i, \underline{x}_i, a_i, r_i, v_i, C_i\}$, where C_i indicates to which cluster a node V_i belongs to. We denote the corresponding partition function by p^C, hence $p^C(V_i) = C_i$. This labelling

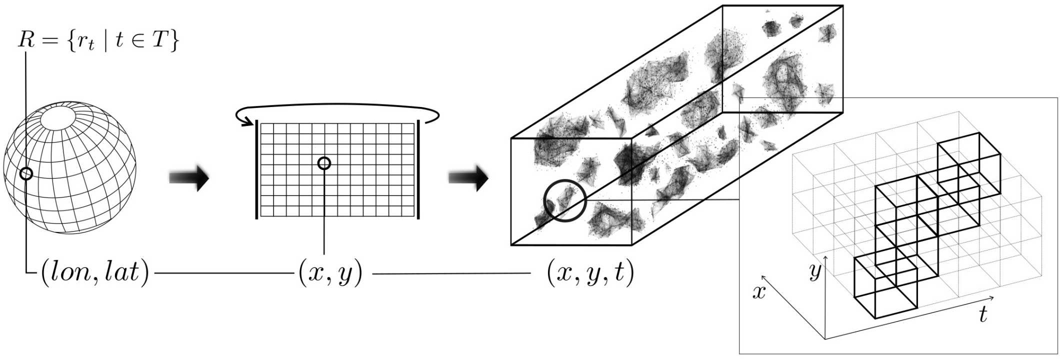


Figure 3.2.: Sketch of spatio-temporal cluster detection. In order to partition the extreme rainfall events into clusters, we first enumerate the given longitude, latitude and time coordinates, allowing us to associate every event with discrete space-time coordinates, $(lon_i, lat_i, t_i) \leftrightarrow (x_i, y_i, t_i)$. Each event can have up to 26 possible neighbors (8 neighbors in the time slice of the measurement, t_i, and 9 neighbors in each of the time slices $t_i - 1$ and $t_i + 1$). Linking all pairs of neighboring extreme rainfall events, we find the clusters by identifying them as the connected components of the graph $G = (V, E)$, where the the set of nodes V is comprised of the events themselves, and the set of edges E indicates whether events are neighbors on the space-time grid. An exemplary cluster is depicted on the very right.

induces a partition of the graph $G = (V, E)$ into n^C spatio-temporal clusters V_i^C of the supergraph $G^C = (V^C, E^C)$, with $V_i^C = \{V_j \,|\, j \in \{1, 2, ..., n\} \wedge p^C(V_j) = C_i\}$.

Next, we compute partition-specific features ${}^C F_i^j$ to assign to the supernodes V_i^C, based on the features of the nodes $V_i \in V$. These features and their calculation are summarized in Tab. 3.2(a).

$$\boldsymbol{G^C = (V^C, E^C)}$$

a) $\boldsymbol{V_i^C}$

feature	symbol	type of feature	given by
${}^C F_i^1$	t_i^{min}	starting time	$\min_{j \in S} t_j$
${}^C F_i^2$	t_i^{max}	end time	$\max_{j \in S} t_j$
${}^C F_i^3$	Δt_i	time span	$t_i^{max} - t_i^{min}$
${}^C F_i^4$	v_i^{sum}	total vol. of water precipitated	$\sum_{j \in S} v_j$
${}^C F_i^5$	L_i^{set}	set of geographical labels	$\{L_j \mid j \in S\}$
${}^C F_i^6$	a_i^{sum}	spatial coverage	$\sum_{L_j \in L_i^{set}} A(L_j)$
${}^C F_i^7$	F_i	family membership	linkage clustering
where $S = \{j \mid j \in \{1, 2, ..., n\} \wedge p^C(V_j) = C_i\}$			

b) $\boldsymbol{E_{ij}^C}$

relation	symbol	type of relation	given by
${}^C R_{ij}^1$	dt_{ij}	temporal distance between clusters	$t_j^{min} - t_i^{min}$
${}^C R_{ij}^2$	IC_{ij}	intersection cardinality	$\lvert L_i^{set} \cap L_j^{set} \rvert$
${}^C R_{ij}^3$	IS_{ij}	intersection strength	$\frac{\mathrm{IC}_{ij}}{\min\{\lvert L_i^{set}\rvert, \lvert L_j^{set}\rvert\}}$

Table 3.2.: The features and relations of the rainfall supergraph $G^C = (V^C, E^C)$**. (a)** The features of the supernodes V_i^C, representing spatio-temporal clusters of extreme rainfall measurements. To compute the spatial coverage of a cluster, we map each geographical grid cell to its surface area, $L_i \mapsto A(L_i)$ (see also the type of feature 'surface area' in Tab. 3.1). The type of feature 'family membership' is introduced in Sec. 3.6. **(b)** The relations of the superedges E_{ij}^C.

3.6. Partitioning into Families of Clusters

We now create superedges between the spatio-temporal clusters, in order to find families of clusters that have a strong regional overlap. Applying the following partition-specific connector function will provide the information necessary for this task,

$$m(V_i^C, V_j^C) := E_{ij}^C = \{dt_{ij}, \mathrm{IC}_{ij}, \mathrm{IS}_{ij}\}, \tag{3.3}$$

where $dt_{ij} = t_j^{min} - t_i^{min}$ is the temporal distance between a pair of clusters, $\mathrm{IC}_{ij} = |L_i^{set} \cap L_j^{set}|$ is the intersection cardinality, which is the number of coinciding geographical grid cells, and $\mathrm{IS}_{ij} = \frac{\mathrm{IC}_{ij}}{\min\{|L_i^{set}|,|L_j^{set}|\}} \in [0, 1]$ is the intersection strength, a measure for the spatial overlap of a pair of spatio-temporal clusters. These properties are also summarized in Tab. 3.2(b).

Based on the above measure of spatial overlap between clusters, we now perform an agglomerative, hierarchical clustering of the spatio-temporal clusters into regionally coherent families. We restrict ourselves to the largest $n^c = 40.000$ clusters with respect to their type of feature 'total vol. of water precipitated', since we are only interested in the strongest extreme rainfall clusters in this chapter. We use the UPGMA algorithm (Sokal et al., 1958) on the distance vector $\underline{d} = (d_{ij})_{i,j \in \{1,...,n^c\}, i<j}$, where $d_{ij} = d(V_i^C, V_j^C) = 1 - \mathrm{IS}_{ij}$, such that we get a total of $n^F = 50$ families. We transfer their labels to both the supernodes of G^C and the nodes of G, hence $V_i^C = \{t_i^{min}, t_i^{max}, \Delta t_i, v_i^{sum}, L_i^{set}, a_i^{sum}, F_i\}$ and $V_i = \{L_i, \underline{x}_i, a_i, r_i, v_i, C_i, F_i\}$, where F_i indicates to which family the node V_i belongs to. We denote the corresponding

$$\boldsymbol{G^F = (V^F, E^F)}$$

a) $\boldsymbol{V_i^F}$

feature	symbol	type of feature	given by
${}^F F_i^1$	T_i^{min}	tuple of start times	$(t_j^{min})_{j \in S}$
${}^F F_i^2$	T_i^{max}	tuple of end times	$(t_j^{max})_{j \in S}$
${}^F F_i^3$	ΔT_i	tuple of time spans	$(\Delta t_j)_{j \in S}$
${}^F F_i^4$	${}^F v_i^{sum}$	total vol. of water precipitated	$\sum_{j \in S} v_j^{sum}$
${}^F F_i^5$	${}^F L_i^{set}$	set of geographical locations	$\bigcup_{j \in S} L_j^{set}$
${}^F F_i^6$	${}^F a_i^{sum}$	spatial coverage	$\sum_{L_j \in {}^F L_i^{set}} A(L_j)$

where $S = \{j \,\vert\, j \in \{1, 2, ..., n^c\} \wedge p^F(V_j^C) = F_i\}$

b) $\boldsymbol{E_{ij}^F}$

relation	symbol	type of relation	given by
${}^F R_{ij}^1$	$(\mathrm{dt}_{uv})_{(u,v) \in S}$	tuple of inter- or intra-family temporal distances	E^C
${}^F R_{ij}^2$	$(\mathrm{IC}_{uv})_{(u,v) \in S}$	tuple of inter- or intra-family intersection cardinalities	E^C
${}^F R_{ij}^3$	$(\mathrm{IS}_{uv})_{(u,v) \in S}$	tuple of inter- or intra-family intersection strengths	E^C

where $S = \{(u, v) \mid u, v \in \{1, 2, ..., n^c\} \wedge p^F(V_u^C) = F_i \wedge p^F(V_v^C) = F_j\}$

Table 3.3.: The features and relations of the rainfall supergraph $G^F = (V^F, E^F)$. **(a)** The features of the supernodes V_i^F, representing families of spatio-temporal rainfall clusters. The first three features are simply the aggregated features of the clusters V_i^C. **(b)** The relations of the superedges E_{ij}^F. They are also just the unprocessed, aggregated relations between intra-family $(i = j)$ and inter-family $(i \neq j)$ clusters.

partition function by p^F, hence $p^F(V_i) = F_i$.

Next, we identify each family of spatio-temporal clusters as a supernode of the induced supergraph $G^F = (V^F, E^F)$, where $V_i^F = \{V_j \,\vert\, j \in \{1, 2, ..., n\} \wedge p^F(V_j) = F_i\}$. Note that, if we were to take the entire set of spatio-temporal clusters, and not just the strongest $n^c = 40.000$, this partition would be a further coarse-graining of the partition induced by p^C, $G \leq G^C \leq G^F$. Therefore, we can redistribute the partition-specific information of G^C, in order to compute the features and relations of G^F as stated in Tab. 3.3(a) and (b), respectively.

We could now compute the temporal inter-cluster intervals of intra-family clusters, or measure the temporal similarities between families. Indeed, the information contained in the properties of G^F can easily be mapped onto event time series. We would only need to identify either T_i^{min} or T_i^{max} as the time index set T_i, and choose the corresponding feature ${}^F F_i^j$ [or function of features $f({}^F F_i^1, ..., {}^F F_i^{f_i})$] as the values v^i,

$$m_b : V^F \to X, V_i^F \mapsto m_b(V_i^F) := X_i = \{v_t^i\}_{t \in T_i}. \tag{3.4}$$

However, in this chapter, we refrain from doing any statistical analysis. Instead, we demonstrate in the next section how the above created deep graph allows us to track and visualize the time evolution of extreme rainfall clusters.

3.7. Exploring Families of Extreme Rainfall Clusters over South America

In the following, we restrict ourselves to two families of spatio-temporal extreme event clusters located over the South American continent. The first family is confined to the subtropical domain [roughly between $40°S$ and $20°S$, see Fig. 3.3(a)], while the second is centered over the tropical Amazon region [roughly between $10°S$ and $10°N$, see Fig. 3.4(a)].

The first family [Fig. 3.3(a)] contains spatio-temporal clusters of extreme events which are characterized by a concise propagation pattern from southeastern South America (around $30°S$, $60°W$) northwestward to the eastern slopes of the northern Argentinean and Bolivian Andes [see Fig. 3.3(b) for an example cluster in this family]. These clusters are remarkable from a meteorological point of view, as their direction of propagation appears to be against the low-level wind direction in this region, which is typically from NW to SE (Vera et al., 2006; Marengo et al., 2012). A case study based on infrared satellite images (Anabor et al., 2008) analyzes some of the “upstream propagating” clusters in this family in detail. This study, together with a detailed climatological analysis of these events using the TRMM 3B42 dataset (Boers et al., 2015), reveals that these spatio-temporal clusters are in fact comprised of sequences of Mesoscale Convective Systems (Maddox, 1980; Durkee et al., 2009; Durkee and Mote, 2009), which form successively along the pathway from southeastern South America towards the Central Andes. The synoptic mechanism explaining this phenomenon is based on the interplay of cold frontal systems approaching from the South, a climatological low-pressure system of northwestern Argentina, and low-level atmospheric moisture flow originating from the tropics (Boers et al., 2015): extensive low-pressure systems associated with Rossby wave trains emanating from the southern Pacific Ocean merge with the low-pressure system over northwestern Argentina to produce a saddle point of the isobars. Due to the eastward movement of the Rossby wave train, the configuration of the two low-pressure systems changes such that the saddle point moves from southeastern South America towards the Central Andes. The deformation of winds around this saddle point leads to strong frontogenesis and hence creates favorable conditions for the development of large-scale organized convection, which explains the observed formation of several mesoscale convective systems along the pathway this saddle point takes. Due to the large spatial extents of these rainfall cluster, as well as due the fact that they propagate into high elevations of the Andean orogen, these systems impose substantial risks in form of flash-floods and landslides, with severe consequences for the local populations. Since this pattern is a recurring feature of the South American Climate system, a complex network approach could recently be employed to formulate a simple statistical forecast rule, which predicts more than 60% of extreme rainfall events at the eastern slopes of the Central Andes (Boers et al., 2014).

The second family [Fig. 3.4(a)] we want to show includes spatio-temporal clusters which exhibit equally concise propagation patterns in the tropical parts of South

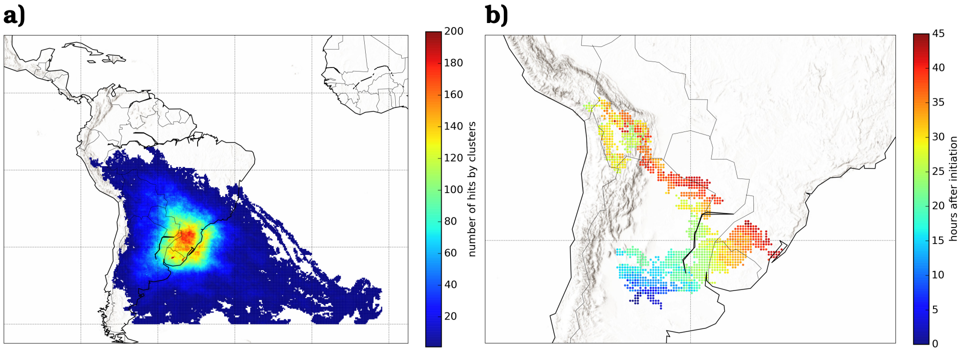


Figure 3.3.: Family of rainfall clusters over subtropical South America. **(a)** The entire family of spatio-temporal clusters over subtropical South America. The colors indicate how often a given grid cell i^L is hit by clusters in this family. This number is given by the number of nodes $n^{FL,i^F i^L}$ in supernode $V^{FL}_{i^F i^L}$ of the intersection partition V^{FL}. Note that the superscript FL indicates that the supernodes $V^{FL}_{i^F i^L}$ arise from intersecting the partitions given by the types of features 'family membership' F and 'geographical label' L. High values over southeastern South America therefore indicate that this is the core region of this family, where most of its clusters pass by in course of their lifetime. **(b)** Exemplary cluster of this family. Each colored grid cell has received at least one event above the 90th percentile belonging to this cluster. The colors indicate the last time (in units of hours) a given grid cell is hit by the cluster, relative to its initiation on February 6, 2011, 18:00 UTC. The temporal evolution of this cluster therefore shows a concise propagation pattern from the Argentinean lowlands across Uruguay toward the eastern slopes of the Central Andes in Bolivia, where the clusters ends on February 8, 2011, 15:00 UTC. This cluster thus lasted for $\Delta t_i = 45h$, and the total sum of water it precipitated was $v_i^{sum} = 4.08 \cdot 10^{10} m^3$, over a total area of $a_i^{sum} = 9.39 \cdot 10^5 km^2$.

America. Similarly to the case described in the previous paragraph, we find several tropical clusters which propagate in the opposite direction of the climatological low-level wind fields. Some of these are initiated at the boundary between tropics and subtropics, move northward along the eastern slopes of the Peruvian Andes, before turning eastward toward the Amazonian lowlands [as for example the cluster shown in Fig. 3.4(b)]. Other instances form just east of the northern Andes, and roughly follow the equator toward the East [as for example the cluster shown in Fig. 3.4(c)]. In view of the above explanations for the first family, we speculate that similar mechanisms leading to the "upstream" propagation of favorable conditions for organized convection are at work in these cases. However, frontal systems do rarely reach these tropical latitudes (Siqueira and Rossow, 2005), and a saddle point similar to the one described above is not present in this case. While Amazonian squall lines, which propagate from the northern Brazilian coast into the continent, have been thoroughly analyzed (Tulich and Kiladis, 2012; Cohen et al., 1995), these organized spatio-temporal clusters moving northward along the tropical Andes and from West to East across the Amazon have – to our knowledge – not yet been studied in the meteorological and climatological literature. We therefore propose these particular spatio-temporal clusters as a promising subject for further research.

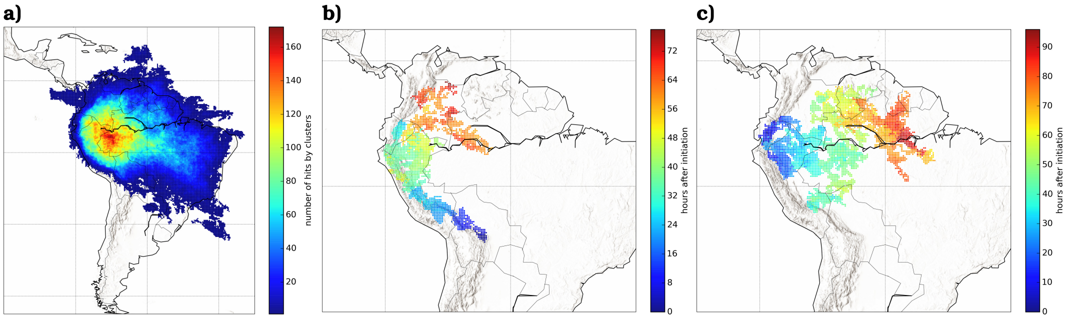


Figure 3.4.: Family of rainfall clusters over tropical South America. (a) The entire family of spatio-temporal clusters over tropical South America. The colors indicate how often a given grid cell i^L is hit by clusters in this family, given by $n^{FL,i^F i^L}$ [see the caption of Fig. 3.3(a)]. High values over the western Amazon therefore indicate that this the core region of this family, where most of its clusters pass by in course of their lifetime. **(b)** First exemplary cluster of this family. Each colored grid cell has received at least one event above the 90th percentile belonging to this cluster. The colors indicate the last time (in units of hours) a given grid cell is hit by the cluster, relative to its initiation on November 4, 2002, 9:00 UTC. The temporal evolution of this cluster therefore shows a concise propagation pattern from central Bolivia northward, along the eastern slopes of the Andes mountain range, before turning west in northern Peru. The cluster ends on November 7, 2002, 15:00 UTC over Colombia and northwestern Brazil, resulting in a total lifetime of $\Delta t_i = 78h$. The total sum of water precipitated by this cluster is $v_i^{sum} = 4.90 \cdot 10^{10} m^3$, covering a total area of $a_i^{sum} = 1.53 \cdot 10^6 km^2$. **(c)** Second exemplary cluster of this family. It initiated on March 20, 2013, 18:00 UTC, at the eastern slopes of the northern Peruvian Andes, and thereafter propagated eastward across the entire Amazon basin, ending on March 24, 2013, 18:00 UTC over northern Brazil. During its lifetime of $\Delta t_i = 96h$, the total sum of water precipitated by this cluster is $v_i^{sum} = 1.17 \cdot 10^{11} m^3$, covering a total area of $a_i^{sum} = 2.61 \cdot 10^6 km^2$.

3.8. Conclusions

In this chapter, we exemplified benefits of the deep graph framework particularly useful in the context of spatio-temporally evolving systems. We did so by conducting an explorative analysis of extreme rainfall events derived from a global, high-resolution satellite product.

First we represented extreme rainfall events as nodes of a graph, whose features indicate their location, time and rainfall rate. By applying suitable connector and selector functions on the set of nodes, we created edges between neighboring events on the spatio-temporal grid prescribed by the resolutions of the data. We then identified cohesive rainfall clusters as the connected components of the graph, allowing us to track and visualize their temporal resolution, and calculate characterizing features: their lifetime, their spatial coverage, and the total volume of water they precipitated. Focussing on the largest 40.000 clusters, we applied another connector function in order to compute the spatial overlap of clusters. This allowed us to create families of clusters with a strong regional overlap.

Finally, we have discussed climatological characteristics of two of these families over the South American continent. The first family, concentrated over the subtropics, was just recently discovered using a rather complicated statistical method. The second, concentrated over tropical South America, has to our knowledge not yet been identified and analyzed in the meteorological literature. We believe these particular clusters would be a promising subject for further research.

The demonstrated approach of tracking and coarse-graining events living in space and time can be easily applied to other spatio-temporal systems. In chapter 5, for instance, we use a similar strategy to characterize fire-cluster burning conditions on different land use types in the Legal Amazon area.

Chapter 4.

The Size Distribution of Extreme Rainfall Clusters around the Globe

4.1. Summary

The scaling behavior of rainfall has been extensively studied both in terms of event magnitudes and in terms of spatial extents of the events. Different heavy-tailed distributions have been proposed as candidates for both instances, but statistically rigorous treatments are rare. Here, we combine the domains of event magnitudes and event area sizes by a spatio-temporal integration of 3-hourly rain rates corresponding to extreme events derived from the quasi-global high-resolution rainfall product TRMM 3B42. A maximum-likelihood evaluation reveals that the distribution of spatio-temporally integrated extreme rainfall cluster sizes over the oceans is best described by a truncated powerlaw, calling into question previous statements about scale-free distributions. The observed sub-powerlaw behavior of the distribution's tail is evaluated with a simple generative model, which indicates that the exponential truncation of an otherwise scale-free spatio-temporal cluster size distribution over the oceans could be explained by the existence of land masses on the globe. This chapter is based on the associated publication P3, and some of the following sections will closely follow parts of this publication.

4.2. Introduction

The spatial and temporal scaling behavior of convection and rainfall has attracted considerable attention in the physical and atmospheric sciences during the past decades. The spatial size distributions of single rainfall events and clouds have been thoroughly analyzed on the basis of various datasets (without considering temporal extents of the events), and these distributions are mostly assumed to be best approximated by a lognormal distribution (e.g. López, 1977; Houze Jr. and Cheng, 1977; Cheng and Houze Jr., 1979; Williams and Houze Jr., 1987). However, recent studies (Mapes and Houze Jr., 1993; Nesbitt et al., 2006) have called this into question, and a powerlaw-type behavior has been proposed as an alternative (e.g. Cahalan and Joseph, 1989; Neggers et al., 2003). Similarly, the scaling properties of rainfall event magnitudes without considering spatial and temporal extents of the events have been studied extensively (e.g. Papalexiou and Koutsoyiannis, 2013;

Serinaldi and Kilsby, 2014), and the number of rainfall events as a function of the event magnitude has been found to exhibit a scale-free range over several orders of magnitude, hinting at similarities to non-equilibrium relaxation processes such as earthquakes or avalanches (Peters et al., 2002; Dickman, 2003; Peters et al., 2010). Indeed, strong empirical evidence has been reported that rainfall might be a real-world example of self-organized criticality (Andrade et al., 1998; Peters and Neelin, 2006; Peters and Neelin, 2009).

However, existing studies investigating the spatial and temporal scaling behavior of rainfall do not perform statistically rigorous comparisons of the proposed powerlaw to alternative heavy-tailed distributions, which may attain shapes that are very hard to distinguish from a true powerlaw (Clauset et al., 2009; Virkar and Clauset, 2014). Whether the relative event magnitude or area size frequencies actually follow a powerlaw distribution has therefore not been rigorously assessed to date. In fact, both the area size and the magnitude distributions have been suggested to exhibit tails which decay faster than that of a powerlaw distribution fitted to the observed values (Peters et al., 2012). Additionally, the scaling characteristics of rainfall have been investigated either in the spatial domain, or in the magnitude domain. For instance, Peters et al. (2012) analyzed the frequency distribution of instantaneous rainfall rates integrated over the spatial extents of the corresponding rainfall cluster - defined there as the set of connected pixels experiencing significant rainfall - but only for single time slices, thus excluding the temporal extents of the events. Also for this quantity, the tail of the frequency distribution decays faster than that of a corresponding powerlaw distribution (Peters et al., 2012).

Here, we analyze the frequency distribution of the total water volume precipitated in spatio-temporally extended extreme rainfall events. The distributional characteristics of the spatio-temporally integrated water volumes, i.e. the total *cluster sizes*, have previously been proposed to follow a scale-free distribution even if this does not hold true for neither the area size, the single-site event magnitudes, nor the spatially integrated event magnitudes (Peters et al., 2012). We use the satellite-derived, gauge-calibrated rainfall dataset TRMM 3B42 (V7), available at 3-hourly temporal resolution on a regular 0.25°-grid covering earth's surface from 50°N to 50°S, and study the spatio-temporal cluster sizes with respect to possible differences over the global oceans and land masses. This is further motivated by the fact that the largest rainfall events at the earth's surface are thunderstorms in the form of Mesoscale Convective Systems and hurricanes (typhoons over the NW Pacific) (Maddox, 1980; Goldenberg, 2001; Zipser et al., 2006), which are - in addition to their spatial sizes - characterized by their outstanding temporal persistence.

We use maximum likelihood estimation (MLE) and maximum likelihood ratio (MLR) comparison tests between several plausible heavy-tailed candidate distributions, and find that the total cluster size distributions over the oceans (land masses) are best described by an exponentially truncated powerlaw (stretched exponential) although they appear to be scale-free over several orders of magnitude. With the help of a simple generative model, we propose the existence of land masses as a possible explanation of the sub-powerlaw behaviour of these distributions.

4.3. The Rainfall Data

As in the previous chapter, we employ the satellite-derived and gauge-calibrated rainfall data product from the Tropical Rainfall Measurement Mission (TRMM 3B42 V7, (Huffman et al., 2007)) with 3-hourly temporal and $0.25° \times 0.25°$ spatial resolutions, for the time period from 1998 to 2014.

Again, we extract extreme rainfall events from the data by considering only those measurements above the 90th percentile of so-called *wet times* (defined as data points with rainfall rates $r \geq 0.1$ mm/h). This wet-time threshold is employed to assure that only data points with significant rainfall are used to compute the distributional characteristics we are interested in here (e.g. Huffman et al., 2007; Scheel et al., 2011; Chen et al., 2013; Zulkafli et al., 2014). The 90th percentile is chosen in agreement with the definition of extreme rainfall events in the IPCC report (Field et al., 2012b) (see Fig. 3.1(a) for the threshold values at each geographical location). This results in $n \approx 2.16 \times 10^8$ extreme events, which we partition into spatio-temporal clusters as described in Sec. 4.4.1.

4.4. Methods

We first recap the concept of spatio-temporal clusters and their sizes from the last chapter. Then, we introduce the different candidate distributions, and the elements of Bayesian parameter inference used to fit candidate distributions to the observed histograms of cluster sizes. Finally, we propose a minimal generative model which reproduces the observed truncated-powerlaw behavior for the cluster size distribution over the global oceans.

4.4.1. Spatiotemporal clusters and their sizes

A spatio-temporal cluster of extreme rainfall is defined as the union of nearest neighbors of extreme rainfall events in the discrete space-time grid prescribed by the resolutions of the TRMM dataset. In order to detect the clusters, we first enumerate the given longitude, latitude and time coordinates. Thereby, we can associate every event with discrete space-time coordinates, $(lon_i, lat_i, t_i) \leftrightarrow (x_i, y_i, t_i)$. Every event can have up to 26 possible neighbors (8 neighbors in the time slice of the measurement, t_i, and 9 neighbors in each of the time slices $t_i - 1$ and $t_i + 1$). Having linked all pairs of neighboring extreme rainfall events, we find the clusters by identifying them as the connected components of the graph $G = (V, E)$, where the set of nodes V is given by the events, and the set of edges E by the links between pairs of neighboring events (see section 3.5 for details of the graph methodology used to determine the spatio-temporal clusters and Fig. 3.2 for an illustration of the clustering scheme). We find a total of $n^C \approx 1.42 \times 10^7$ spatio-temporal clusters. For each of these clusters, we compute the total volume of water precipitated (of dimension: length^3), i.e. the cluster size s of cluster C, by

$$s = \sum_C a_i \cdot r_i \cdot 3h, \tag{4.1}$$

where a_i is the surface area of a rainfall event, given by

$$a_i = (111km)^2 \cdot (0.25)^2 \cdot \cos\left(\frac{2\pi}{360^\circ} \cdot lat_i\right). \tag{4.2}$$

This definition of cluster size follows the definition of "event size" in Peters et al., 2012. Note that an alternative choice of defining the size of clusters would be to simply take the number of events belonging to a cluster. In fact, for the clusters obtained as described above, the total volume of precipitated water is highly correlated to the number of events ($r_{\text{Pearson}} = 0.98$). However, we chose the total volume of water as the metric for cluster sizes, since it is more accurate.

The set of all clusters is then partitioned into three groups: clusters that precipitated mainly above ocean, mainly above land, and a group for the remaining clusters. For this, we first use a land-ocean mask to classify every rainfall event as an ocean or a land event. We then decide that a cluster belongs to the ocean (land) group if more than 90% of its constituent events are ocean (land) events, leading to $n^C_{\text{ocean}} \approx 8.68 \times 10^6$ ($n^C_{\text{land}} \approx 5.20 \times 10^6$) clusters. The remaining cluster are attributed to the mixed group, with $n^C_{\text{mixed}} \approx 3.55 \times 10^5$ elements. Note that the results presented below are insensitive to changing this parameter, e.g., from 90% to 100%.

4.4.2. Estimation of the distributions of spatio-temporal extreme rainfall clusters

For all clusters combined, the ocean and land groups separated, and the generative model introduced in the next section, we fit a set of candidate distributions to the cluster size distributions. These candidates include a powerlaw (PL), a truncated powerlaw (TPL), a stretched exponential (SEXP), and a log-normal (LN), given by

$$\text{PL: } f(x) = \frac{(\alpha-1)}{x_{min}}\left(\frac{x}{x_{min}}\right)^{-\alpha} \tag{4.3}$$

$$\text{TPL: } f(x) = \frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha, \lambda x_{min})} x^{-\alpha} e^{-\lambda x} \tag{4.4}$$

$$\text{SEXP: } f(x) = \beta\lambda x^{\beta-1} e^{-\lambda(x^\beta - x^\beta_{min})} \tag{4.5}$$

$$\text{LN: } f(x) = \sqrt{\frac{2}{\pi\sigma^2}}\left[\text{erfc}\left(\frac{\ln x_{min} - \mu}{\sqrt{2}\sigma}\right)\right]^{-1} \frac{1}{x}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \tag{4.6}$$

In a first step, optimal functional forms for these distributions with respect to the observed cluster sizes are determined by MLE (Clauset et al., 2009): For each proposed candidate ρ, the likelihood of its parameters $\mathcal{P}$, given the set of observed cluster size values $\mathcal{C}$ is maximized. Under the assumption of flat priors $P(\rho, \mathcal{P})$, Bayes' Theorem assures that the parameters determined via this optimization are the most likely parameters given the observed data $\mathcal{C}$:

$$P(\rho, \mathcal{P}|\mathcal{C}) = \frac{P(\mathcal{C}|\rho, \mathcal{P})\, P(\rho, \mathcal{P})}{P(\mathcal{C})} \; , \tag{4.7}$$

where $P(\mathcal{C})$ is unknown and in practical terms impossible to compute. The likelihood of the parameters $\mathcal{P}$ given the data $\mathcal{C}$ is defined as $\mathcal{L}_{\mathcal{C}}(\rho, \mathcal{P}) \equiv P(\mathcal{C}|\rho, \mathcal{P})$, and assuming that $P(\rho, \mathcal{P})$ is non-informative (i.e., a flat distribution), we have $P(\rho, \mathcal{P}|\mathcal{C}) \propto \mathcal{L}_{\mathcal{C}}(\rho, \mathcal{P})$. We are thus left with the following optimization problem:

$$\mathcal{P}^* = \arg\max_{\mathcal{P}} \mathcal{L}_{\mathcal{C}}(\rho, \mathcal{P}) = \arg\max_{\mathcal{P}} \prod_{i=1}^{n^C} \rho(s_i; \mathcal{P}) \; , \tag{4.8}$$

where n^C denotes the number of clusters and s_i their respective spatio-temporally integrated size. The optimal parameters thus determined for the four candidate distributions are listed in the legends of Figs. 4.1(a)-(c).

The likelihood of each candidate, evaluated with the respective MLE-optimal parameters, is then compared by means of a MLR comparison test. The Neyman-Pearson lemma assures that this is the most efficient statistical test possible to compare between two candidate distributions (Neyman and Pearson, 1933). Setting $L_{\mathcal{C}}(\rho) = \max_{\mathcal{P}} \mathcal{L}_{\mathcal{C}}(\rho, \mathcal{P})$, we compute for two candidates ρ_1 and ρ_2 the log-ratio

$$\mathcal{R}_{\mathcal{C}}(\rho_1, \rho_2) = \log \frac{L_{\mathcal{C}}(\rho_1)}{L_{\mathcal{C}}(\rho_2)} \; . \tag{4.9}$$

If $R_{\mathcal{C}}(\rho_1, \rho_2) > 0$ (< 0), we conclude that ρ_1 is a more (less) likely model of the observed cluster size distribution than ρ_2. A test of statistical significance for the values of $\mathcal{R}$ can be derived from the central limit theorem (see Clauset et al. (2009) for details, in particular for cases where the two distributions to be compared are nested versions of each other).

4.4.3. Generative model for spatio-temporal cluster sizes

Here, we introduce a generative model to test the hypothesis that the sub-powerlaw behaviour of observed spatio-temporal extreme rainfall cluster size distributions is due to the existence of land masses on earth. The model is motivated by the assumption of a scale-free (i.e., PL) distribution of rainfall cluster sizes, whose truncation is caused by the fact that hurricanes end prematurely as soon as they hit the coast. The model is designed as follows: A synthetical storm with a lifetime drawn from a PL distribution is placed on a random pixel of a cage consisting of 1000 × 1000 pixels. The temporal evolution of the storm is prescribed by randomly selecting a neighboring pixel at each time step, where the selection probabilities depend on the previous direction of movement. Moving in the same direction as before has a higher probability ($p = 0.95$) than turning left ($p = 0.04$) or right ($p = 0.01$), and the probability to move backwards is zero. The storm keeps moving either until its predetermined (PL-distributed) lifetime ends, or until it hits the boundary of the cage, which immediately ends the lifetime of the storm. With each time step, the storm grows by one unit, hence its "size" is proportional to its lifetime. The experiment is simulated with a large number of storms (10.000), for each of which we record its effective lifetime. The resulting distribution of lifetimes is shown in Fig. 4.1(d) and discussed in Sec. 4.5.

The inertia of the storm's movement scheme imposed by the dynamic selection probabilities ensures more realistic storm tracks: First, a storm is very unlikely to move backwards (hence $p = 0$ for turning backwards). Second, the symmetry of moving left or right is broken in order to mimic the influence of the Coriolis force. However, we found that changing the selection probabilities, or removing the inertia entirely ($p = 0.25$ for each direction) does not lead to qualitatively different lifetime distributions.

The assumptions behind this generative mechanism is that large storms such as hurricanes typically initiate over the ocean, and that their lifetimes over the oceans would be PL-distributed due to the abundant energy provided by the ocean. However, most large storms eventually hit the coast, where the lack of available energy for their persistence causes them to die out (Whitaker and Davis, 1994; Briegel and Frank, 1997; Raymond and Sessions, 2007; Nolan et al., 2007). The lifetimes can be approximately taken to be proportional to the sizes of the storms in terms of total precipitated water. Therefore, this model is suitable to investigate how the boundaries of the oceans (i.e., the coasts) would impact a cluster size distribution prescribed as a PL.

4.5. Results

The observed spatio-temporally integrated extreme rainfall cluster sizes for all clusters (Fig. 4.1(a)) as well as for clusters over the global oceans (Fig. 4.1(b)) extend beyond 10^3km^3 of precipitated water. In contrast, the cluster size distribution over the global land masses only reaches sizes up to 10^2km^3 (Fig. 4.1(c)). Out of the four distributions proposed as candidates for the cluster sizes (PL, TPL, SEXP, and LN), the SEXP wins the MLR comparison test (see Sec. 4.4.2) for the combination of all clusters. The TPL wins for the subset of clusters over the oceans. Note that visually, the SEXP may appear to be the better model. This is, however, only caused by the fact that the log-log representation of the probability density functions emphasizes the tail of distributions rather than the comparably smaller cluster sizes, which have higher probability weights by several orders of magnitude. For clusters over the global land masses, the SEXP is again the most likely candidate distribution. For comparison, the generative model of synthetical storms, with PL-distributed lifetimes that are put into a cage with absorbing boundaries, leads to a TPL distribution of the effective lifetimes as the most likely candidate (Tab. 4.1).

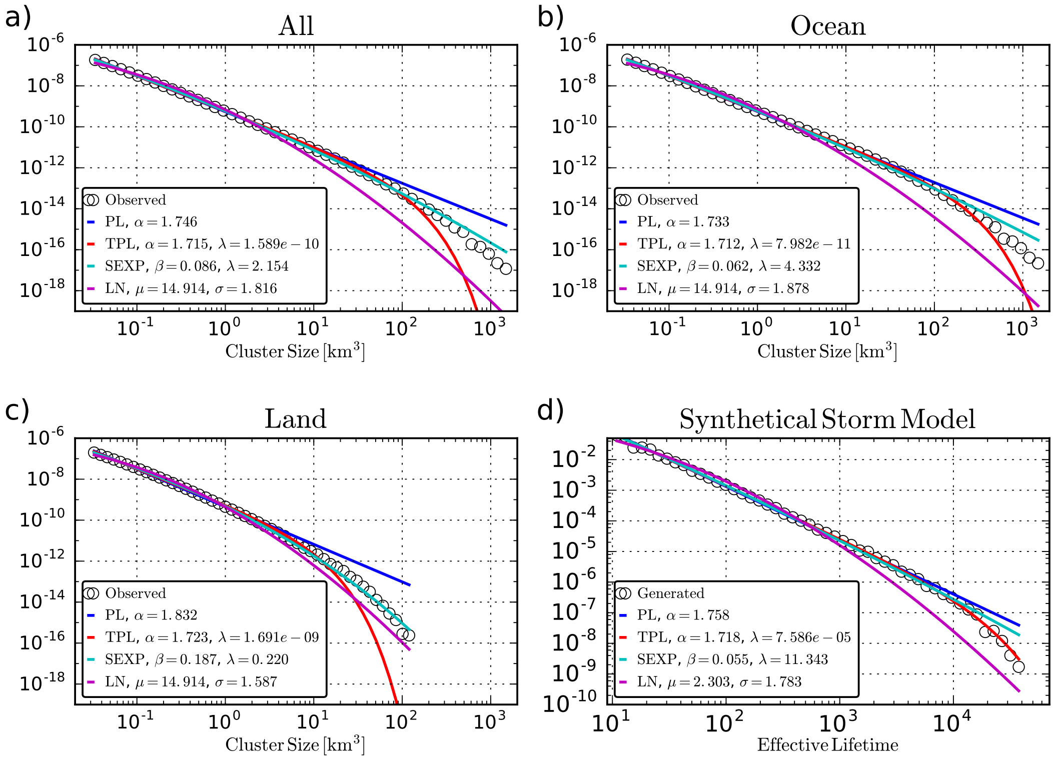


Figure 4.1.: Histograms of the observed spatio-temporal extreme rainfall cluster sizes and the effective lifetimes of the generative model introduced in Sec. 4.4.3 (circles), as well as the corresponding MLE-optimized fits of the proposed candidate distributions (lines). In all panels, blue lines indicate the optimal powerlaw (PL) fits, red lines the truncated powerlaw (TPL) fits, cyan lines the stretched exponential (SEXP) fits and magenta lines the log-normal (LN) fits. The respective optimal parameters of the MLE fits are stated in the legends. **(a)** Histogram of observed cluster sizes for all clusters combined ($n^C \approx 1.42 \cdot 10^7$) and MLE-fitted cluster size distributions. **(b)** Histogram of observed cluster sizes for the subset of ocean clusters ($n^C_{\text{ocean}} \approx 8.68 \cdot 10^6$), and MLE-fitted cluster size distributions for this subset. **(c)** Histogram of observed cluster sizes for the subset of land clusters ($n^C_{\text{land}} \approx 5.20 \cdot 10^6$), and MLE-fitted cluster size distributions for this subset. **(d)** Histogram of effective lifetimes obtained from the generative model introduced in Sec. 4.4.3, and corresponding MLE-fitted lifetime distributions. Note that the units are arbitrary in this case.

	All	Ocean	Land	Model
$R_{\mathcal{C}}(\rho_{\mathrm{TPL}}, \rho_{\mathrm{PL}})$	14778	5366	19865	339
$R_{\mathcal{C}}(\rho_{\mathrm{SEXP}}, \rho_{\mathrm{PL}})$	22252	3751	24404	191
$R_{\mathcal{C}}(\rho_{\mathrm{SEXP}}, \rho_{\mathrm{TPL}})$	7474	-1615	4539	-148
$R_{\mathcal{C}}(\rho_{\mathrm{LN}}, \rho_{\mathrm{PL}})$	-154656	-129157	-8909	-4038
$R_{\mathcal{C}}(\rho_{\mathrm{LN}}, \rho_{\mathrm{TPL}})$	-169435	-134523	-28774	-4377
$R_{\mathcal{C}}(\rho_{\mathrm{LN}}, \rho_{\mathrm{SEXP}})$	-176908	-132908	-33312	-4229

Table 4.1.: MLR test results for the comparisons between candidate distributions for cluster sizes. MLR test results for the comparisons between all considered candidate distributions (PL, TPL, SEXP and LN) for the observed cluster size distributions ("All" clusters, "Ocean" clusters and "Land" clusters) and the lifetime distribution of the synthetical storm model ("Model"). We recall from Sec. 4.4.2 that a distribution ρ_1 is more (less) likely than another distribution ρ_2 if $R_{\mathcal{C}}(\rho_1, \rho_2) > 0$ (< 0). All corresponding p-values are smaller than 10^{-30}.

A visual comparison of all four histograms with their respective optimal PL-fits (a straight line in the log-log-plots of Fig. 4.1) may suggest a scale-free distribution at least over several orders of magnitude (approximately: for all clusters up to $20\,\mathrm{km}^3$, for ocean clusters up to $40\,\mathrm{km}^3$, for land clusters up to $3\,\mathrm{km}^3$ and for the generative model up to 4000 units). However, the PL does not win the MLR comparison even when only considering cluster sizes and lifetimes below these values. Towards larger cluster sizes, all observed histograms clearly show dampened tails as compared to the PL, albeit to different extents.

To get a visual impression of clusters at the tail-end of the distributions, we show the time evolution of the largest cluster observed over land, as well as the largest clusters over the Atlantic, Pacific, and Indian Ocean (Fig. 4.2). The partitioning of extreme rainfall events into clusters of (spatio-temporally) connected neighboring pixels can lead to clusters with very clear trajectories (Figs. 4.2(a) and (c)), but also to rather scattered clusters without a clear propagation direction (Figs. 4.2(b) and (d)). Amongst the largest clusters over the oceans we found - by visual inspection - a large number of hurricanes (typhoons), such as the one over the Atlantic (Fig. 4.2(a)). The largest land cluster in the data (Fig. 4.2(c)) has recently been discussed by Boers et al. (2014), Boers et al. (2015), and Traxl et al. (2016).

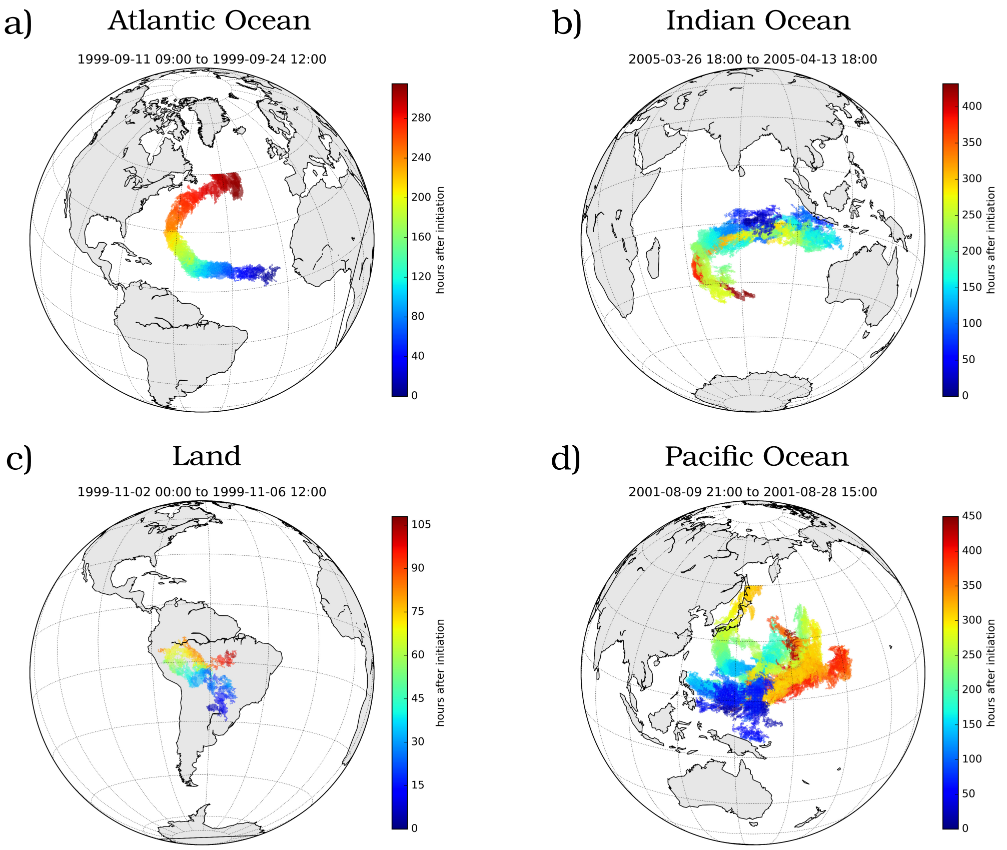


Figure 4.2.: Largest clusters of the land and different ocean groups. For each panel, each colored grid cell has received at least one event above the 90th percentile belonging to the respective cluster. The colors indicate the last time (in units of hours) a given grid cell is hit by the cluster, relative to its initiation time (see titles of the respective panels). **(a)** The largest cluster that precipitated above the Atlantic ocean. **(b)** The largest cluster that precipitated above the Indian ocean. **(c)** The largest cluster that precipitated above land. **(d)** The largest cluster that precipitated above the Pacific ocean.

4.6. Discussion

Our results indicate that none of the observed spatio-temporally integrated extreme rainfall cluster size distributions are described well by a PL. This is in contrast to propositions of scale-free distributions in the literature (Cahalan and Joseph, 1989; Peters et al., 2002; Dickman, 2003; Neggers et al., 2003), but also corroborates earlier results obtained for cluster sizes defined by either a temporal or a spatial integration of rain rates (Peters et al., 2012). Although the size distributions of rainfall clusters over the lands and oceans are visually similar to the PL-fits over several orders of magnitude, both of them decay faster than the corresponding PL at the tails of the distributions. The MLR test results (SEXP for all clusters combined, TPL for ocean clusters, and SEXP for land clusters) suggest that the influence of the subsets of land and mixed clusters on the size distribution of all clusters combined leads to a SEXP also for all clusters. However, clusters larger than 10^2km^3 occur almost exclusively over the oceans, and this is also the domain where the dampening of the distribution as compared to a corresponding PL becomes most apparent (Fig. 4.1(a)).

The proposed generative model (Sec. 4.4.3) tries to explain the observed behavior by postulating that the ocean cluster sizes would indeed follow a scale-free (i.e., PL) distribution on a planet without land masses, and that the exponential truncation occurs primarily due to the fact that hurricanes end prematurely as soon as they hit the coast. Physically, this hypothesis can be motivated as follows: While the specific mechanisms of cyclogenesis are still not entirely understood, it is clear that cyclones can only maintain themselves over the oceans, under the conditions (among others) of sufficiently warm sea surface temperatures and strong, moist convection, which guarantee that enough latent heat can be released to the atmosphere to fuel the storms. Shortly after their landfall, the cyclones die because these conditions are no longer fulfilled (Whitaker and Davis, 1994; Briegel and Frank, 1997; Raymond and Sessions, 2007; Nolan et al., 2007). In the generative model, storms with a relatively longer lifetime have a higher probability of hitting a border of the cage at some point in their life span, which leads to the observed truncation of the effective lifetime distribution (Fig. 4.1(d)). The analogy to the ocean cluster size distribution is that it would in fact also follow a powerlaw, if there were no land masses on the planet. Hence, rainfall clusters could freely move, like the synthetical storms without a cage, on this imaginary aqua planet, with their size distribution predetermined by a powerlaw. The fact that there are land masses on earth would then lead to an exponential truncation of the PL distribution (Fig. 4.1(b)), as it is the case for the synthetical storms of the generative model (Fig. 4.1(d)).

We note that the proposed hypotheses corresponds to a very simplified view of the complex physical processes involved in the formation and maintenance of cyclones. We do not propose that the only relevant physics behind the exponential truncation of extreme rainfall cluster size distributions are given by the stated hypotheses. It is, however, remarkable that the generative model is – despite its simplicity – capable of reproducing the observed statistical characteristics of rainfall cluster sizes over the oceans. From a purely statistical point of view, our hypothesis can hence not be rejected.

4.7. Conclusion

Based on the high-resolution TRMM 3B42 dataset, we have provided statistically rigorous evidence that the spatio-temporally integrated size distribution of extreme rainfall clusters does not - as previously suggested - follow a powerlaw. Instead, we find that the size distribution of rainfall clusters over the oceans (land masses) is best approximated by an exponentially truncated powerlaw (stretched exponential), leading to a stretched exponential as the most likely candidate distribution for all clusters combined (land and ocean together). We hypothesize that the size distribution of extreme rainfall clusters over the oceans could, in principle, follow a scale-free distribution on a planet without land masses, and that the exponential truncation of the observed distribution is caused by the presence of land masses. Physically, this is motivated by the fact that the conditions for cyclogenesis are not met over land. To test this hypothesis, we proposed a simple generative model of synthetic storms with powerlaw distributed lifetimes, evolving in a finite spatial area with absorbing boundaries. This simple model reproduces the exponentially truncated powerlaw observed for extreme rainfall clusters over the oceans, indicating that the proposed hypothesis suffices to explain the distributional characteristics discovered here.

Chapter 5.

Characterizing Fire-Cluster Burning Conditions on different Land Use Types

5.1. Summary

The Amazon rainforest has a major influence on carbon storage and climate, and plays a key role in reducing pollutant levels on a regional and global scale. Unfortunately, it is becoming increasingly vulnerable to catastrophic fires due to a combination of droughts, climate change and human activities such as conversions of natural vegetation into pasture and agricultural fields and deforestation in general. We advance the understanding to what extent different land use types influence fire occurrence in the Amazonian ecosystem, which is particularly relevant for its conservation. Based on a combination of two high-resolution satellite products - maps of fire-affected areas and land cover maps showing a detailed land use classification - the deep graph framework is employed to identify spatio-temporal fire clusters in the Legal Amazon region, and their land use specific burning conditions are characterized statistically. For each identified cluster, we generate a set of features: its size; lifetime; dominant land use type (given by the most frequent land use type within the spatial domain of the cluster); and the land use type of the location where the fire started. We find that the distributions of diameters and lifetimes are dominated by clusters that occur in savannah-type ecosystems, not only in terms of frequency, but also in largest sizes and longest lifetimes. This is followed by forest-fires, fires on pasture fields and fires on agricultural fields, which show a consistent decrease in frequency and slopes. The least frequent, smallest and shortest clusters occur on secondary vegetation and deforested areas. By means of likelihood-ratio tests we find that all diameter and lifetime distributions exhibit heavy tails, i.e. their tails are not exponentially bounded. With respect to the originating land use type(s) of clusters, we found that 19% of all identified fire clusters have “Pasture”, 5% have “Agriculture”, and 2% have “Deforested” in their set of land use types measured on the first satellite pass. Finally, we derive probabilistic classifiers of fire clusters into dominant land use types, based on different combinations of their features. Considering either diameters or lifetimes of clusters separately, we find that the probability of finding fire clusters other than savannah-type clusters rapidly declines with increasing diameters (lifetimes). Overall, the best bet for any given cluster’s diameter (lifetime) is to classify it as savannah-type. We increase the separability of classes by taking into account both a cluster’s lifetime

and its diameter, and find that for certain combinations, clusters occurring on pasture fields and forests are more likely to be found than savannah-type clusters. The classifiers also provide for insightful visualizations of the burning characteristics for the different kinds of clusters. Finally, we discuss ideas how to improve the classifiers' predictive power. This chapter is part of publication P4, which is in preparation.

5.2. Introduction

In this chapter, we employ the deep graph framework to identify spatio-temporal fire clusters in the Legal Amazon - a region in the Amazon basin covering more than 5 million square kilometers - and then characterize their land use specific burning conditions.

Investigating the land use specific burning conditions of fires in the Amazon basin is of particular interest. First, because the Amazon rainforest has a major influence on carbon storage and climate, and plays a key role in reducing pollutant levels on a regional and global scale (Laurance, 1999; Cochrane and Laurance, 2008; Liu et al., 2014). Without the Amazon rainforest, the greenhouse effect would likely be more pronounced, and climate change may possibly get worse in the future. Second, because a vast majority of burning events in the Amazon rainforest result from anthropogenic activities, whereas natural fire occurrences are extremely rare (Cochrane, 2003; Asner et al., 2005; Laurance et al., 2001). There is an intensive, deliberate use of fire to convert natural vegetation into pasture and agricultural fields (Morton et al., 2006; Armenteras and Retana, 2012; Davidson et al., 2012), to maintain deforested areas (Cochrane et al., 1999; Barona et al., 2010), and to re-new the grass of pasture fields for the cattle of farmers. It has been shown that escaping fires from managed pastures and agricultural lands significantly contribute to forest fires (Cano-Crespo et al., 2015). Therefore, advancing our understanding to what extent different land use types influence fire occurrences in the Amazonian forest is particularly relevant for its conservation.

Here, we combine two different, high-resolution satellite products - maps of fire-affected areas and land cover maps showing a detailed land use classification - into a graph representation. Nodes of the graph represent pixels within the Legal Amazon region that are affected by fire. Their features indicate location and time of the fire, combined with the land use type it occurred on. Similarly to chapter 3 and chapter 4, edges between nodes are created upon spatio-temporal proximity, expressing the fact that fires at close-by locations might be related in the sense that one fire was caused by the other through propagation. We identify spatio-temporal fire clusters as the connected components of the graph, and generate a set of features for each of them: the size; the lifetime; the dominant land use type; and the originating land use type(s) of a cluster.

Based on these characteristics, we first investigate the size and lifetime distributions of fire clusters, resolved with respect to their different dominant land use types. Then, we inspect the originating land use types for the different types of clusters. Finally, we derive probabilistic classifiers of fire clusters into dominant land use types, based on different combinations of their features. These classifiers also provide for insightful visualizations of the burning characteristics for the different kinds of clusters.

5.3. The Active Fire and Land Use Type Data

5.3.1. Fire

The Moderate Resolution Imaging Spectroradiometer (MODIS) on board the polar-orbiting Terra and Aqua satellites maps fire-affected areas since 2000. We employ the MODIS collection 5 global fire location product (MCD14ML) developed by Giglio et al. (2003) in the Legal Amazon area, for the years 2008 and 2010 (for other years, no land use type data is available). Every 1 km spatial resolution active fire (AF) observation holds information about the location and time when it was detected by the sensors.

Fire detection is performed using a contextual algorithm that exploits the strong emission of mid-infrared radiation from fires. The algorithm examines each pixel of the MODIS swath, and ultimately assigns one of the following classes to each of them: missing data, cloud, water, non-fire, fire, or unknown. In this thesis, we are only considering pixels classified as fire with a confidence larger than 60%.

5.3.2. Land Use Type

Land cover maps of the Legal Amazon were produced by the TerraClass project showing a very detailed land use classification at 30m resolution with data generated from the interpretation of Landsat Thematic Mapper 5 images (Almeida et al., 2009).

The maps include 15 disjoint land use classes in 2008 and 16 in 2010 (a reforestation class was introduced). We employ the six classes listed below and aggregate the rest to a “Others” class:

- *Nonforest* covers natural vegetation with characteristics of savannah-type ecosystems: cerrado, campinas or campinaranas.

- *Forest* refers to native tree vegetation with no or little disturbance and continuous canopy.

- *Pasture* refers to areas currently in use for grazing where there is a predominance of herbaceous vegetation and 90-100% grass coverage.

- *Agriculture* refers to large areas with predominance of annual crops that use certified seeds, pesticides and mechanization, among others.

- *Secondary vegetation* refers to areas that after a total removal of the tree vegetation are in an advanced stage of shrub and/or tree vegetation.

- *Deforestation* refers to deforested areas.

- *Others* includes water bodies and fluvial beaches, urban areas, mines, sand bars, rock outcrops, bare soil and unclassified areas.

5.4. Representing the Data as Nodes of a Graph

We process the data of the two satellite products such that each active fire measurement with a confidence larger than 60% corresponds to a node V_i of the graph $G = (V, E)$. There is a total of $n = 393.474$ nodes, each with the same set of types of features: spatial coordinates given by a latitude and longitude pair, the date and time of the measurement, and the land use type the fire occurred on. These features are summarized in Tab. 5.1(a).

The spatial distribution of active fires in 2008 and 2010 along with their corresponding land use types is illustrated in Fig. 5.1. The monthly distribution of events in 2008 and 2010, as well as the distribution of events across land use types is depicted in Fig. 5.2.

$$\boldsymbol{G = (V, E)}$$

a)

$\boldsymbol{V_i}$

feature	symbol	type of feature	given by
F_i^1	lat_i	latitude	the data
F_i^2	lon_i	longitude	the data
F_i^3	t_i	time	the data
F_i^4	lut_i	land use type	the data
F_i^5	C_i	cluster membership	connected components

b)

$\boldsymbol{E_{ij}}$

relations	condition	symbol	type of relation	given by
R_{ij}^1	if $\lvert dt_{ij}\rvert \leq$ TT and $gcd_{ij} \leq$ ST	dt_{ij}	temporal distance	$\lvert t_j - t_i\rvert$
R_{ij}^2	if $\lvert dt_{ij}\rvert \leq$ TT and $gcd_{ij} \leq$ ST	gcd_{ij}	great-circle distance	see Eq. (5.2)
$\emptyset$	if $\lvert dt_{ij}\rvert >$ TT or $gcd_{ij} >$ ST			

Table 5.1.: The features and relations of the graph $G = (V, E)$. **(a)** The features of the nodes V_i, representing active fire measurements. The type of feature 'cluster membership' is introduced in Sec. 5.5. **(b)** The relations of the edges E_{ij}, representing the spatio-temporal distance between active fire measurements. An edge only exists, if the condition stated in the table is fulfilled.

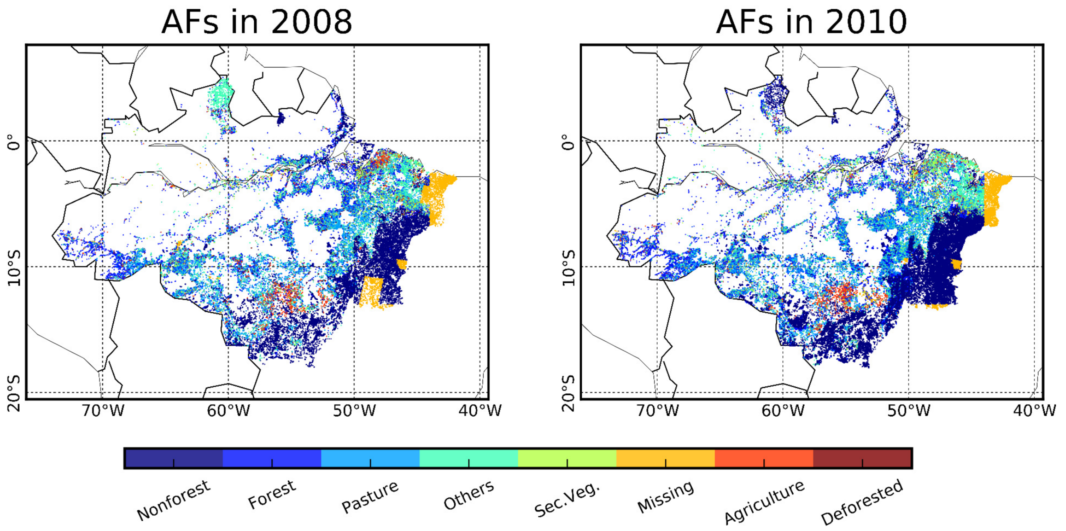


Figure 5.1.: Spatial distribution of active fire measurements in 2008 and 2010. The spatial distribution of active fire (AF) measurements for the years 2008 and 2010. Every measurement of a fire corresponds to a colored pixel on the maps. The colors indicate the land use type the fires occurred on. Orange pixels ("Missing") correspond to unknown land use types, either because they are outside the Legal Amazon region (eastern Maranhao or southern Tocantins), or because there were errors in the measurement process.

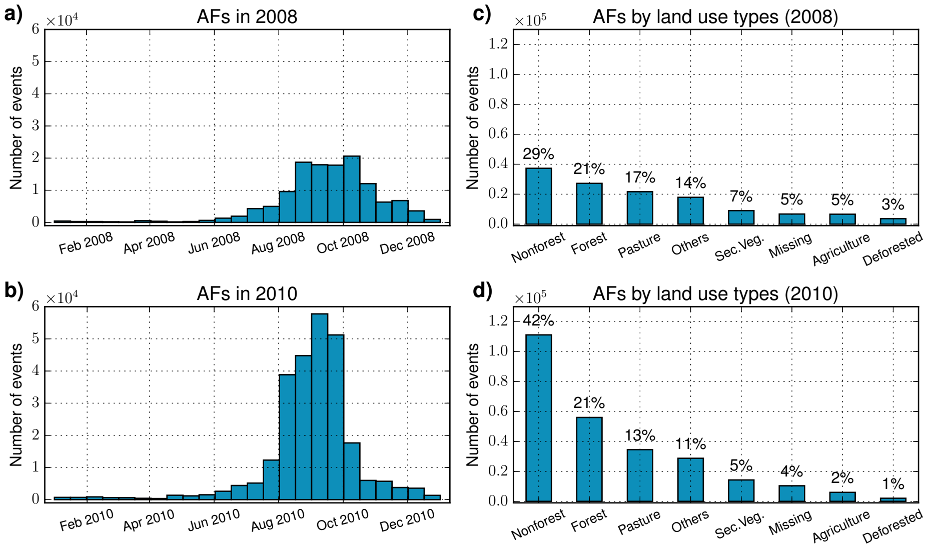


Figure 5.2.: Distribution of active fires across months and land use types. The distribution of active fire measurements across months [(a)-(b)] and across land use types [(c)-(d)] for the years 2008 and 2010. **(a)-(b)** A pronounced peak is visible for the occurrence of fires in the dry season, centered around September each year. Almost twice as many fires have been observed in 2010 compared to 2008 (263.289 versus 130.185, respectively), due to the extreme drought in 2010 Lewis et al., 2011. **(c)-(d)** The numbers above the bars represent the proportion of fires occurring on the different land use types, within each year.

5.5. Partitioning into Spatio-temporal Fire Clusters

To understand the burning conditions of fires - such as typical spatial extents and lifetimes - we need a notion of spatio-temporal clusters of fire. Here, we follow a similar procedure as laid out in chapter 3, a clustering based on the spatial and temporal proximity of active fires measured by the Terra and Aqua satellites. This procedure is described in the following.

First, we compute the spatial and temporal distances between pairs of active fire measurements by defining the following connector function on the set of nodes,

$$m(V_i, V_j) := E_{ij} = \{dt_{ij}, gcd_{ij}\}, \tag{5.1}$$

where $dt_{ij} = |t_j - t_i|$ is the temporal distance, and gcd_{ij} the great-circle distance between a pair of fire measurements. The great-circle distance is given by

$$gcd_{ij} = R \cdot \arccos\left(\sin\phi_i \cdot \sin\phi_j + \cos\phi_i \cdot \cos\phi_j \cdot \cos(\lambda_j - \lambda_i)\right), \tag{5.2}$$

where $\phi_x = \frac{\pi}{360} \cdot lat_x$, $\lambda_x = \frac{\pi}{360} \cdot lon_x$ and $R = 6.371$ km (the average radius of earth). The set of possible all edges is therefore given by $E' = \{E_{ij} \,|\, i, j \in \{1, 2, ..., n\}\}$.

Next, we need to find appropriate spatial and temporal thresholds. They represent the upper bounds on the spatio-temporal distances for which we deem two events related, in the sense that one was caused by the other through propagation. Unfortunately, the fire measurements do not come on an equidistant spatio-temporal grid, making it more difficult to find appropriate thresholds as compared to the rainfall data analyzed in chapter 3 and chapter 4. We first have to look at the spatial and temporal resolutions of the satellite measurements. In Fig. 5.3(a), the distribution of fire measurements across the hours of the day is depicted. The Terra and Aqua satellites each perform two measurements a day, which is clearly visible by the four peaks in the histogram (at around 02:00 am, 05:00 am, 14:00 pm and 17:00 pm). The temporal inhomogeneity of the measurements is also reflected in the distribution of temporal distances of the edges, depicted in Fig. 5.3(b). To allow all events to link to any other event in the next satellite pass, the minimal temporal threshold is approximately 14 hours, $TT = 14$ h. The distribution of spatial distances is illustrated in Fig. 5.3(c). It reveals a peak at approximately one kilometer spatial distance, which is the stated resolution of the satellites. For the following analysis, we choose a spatial distance threshold of two kilometers, $ST = 2$ km. So far, the spatio-temporal thresholds have been validated by inspection of the later derived cluster-size distributions, for which they provide plausible results.

Given the temporal threshold of $TT = 14$ hours and the spatial threshold of $ST = 2$ kilometers, we can compute the spatio-temporal fire clusters by identifying them as the connected components of the graph $G = (V, E)$, where E is given by applying the selector

$$s(E_{ij}) := \begin{cases} E_{ij} = \{dt_{ij}, gcd_{ij}\} & \text{if } |dt_{ij}| \leq TT \wedge gcd_{ij} \leq ST \wedge i \neq j \\ \varnothing & \text{else} \end{cases} \tag{5.3}$$

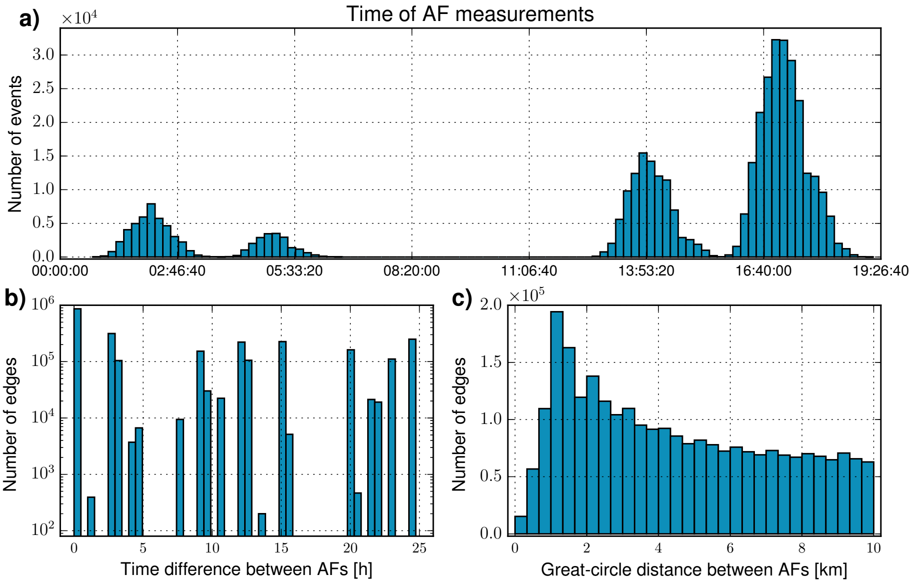


Figure 5.3.: Distribution of fire measurements across hours of the day and edge weight distributions. (a) Histogram of the active fire measurements across the hours of the day (all events of 2008 and 2010 are included). Four peaks are clearly visible, which correspond to the two satellite passes of each the Terra and Aqua satellites. **(b)** Histogram of the temporal distances between active fire measurements (only shown for time differences up to 25 hours). The temporal inhomogeneity of the fire measurements is clearly visible. **(c)** Histogram of the great-circle distances between active fire measurements (only shown for distances up to 10 kilometers).

$$\boldsymbol{G^C = (V^C, E^C)}$$

	feature	symbol	type of feature	given by
$\boldsymbol{V_i^C}$	${}^CF_i^1$	t_i^{min}	starting time	$\min_{j \in S} t_j$
	${}^CF_i^2$	t_i^{max}	end time	$\max_{j \in S} t_j$
	${}^CF_i^3$	Δt_i	lifetime	$t_i^{max} - t_i^{min}$
	${}^CF_i^4$	d_i	cluster diameter	$\max_{j,k \in S} gcd_{jk}$
	${}^CF_i^5$	lat_i^{mean}	mean latitude	$\frac{1}{\vert S \vert} \sum_{j \in S} lat_j$
	${}^CF_i^6$	lon_i^{mean}	mean longitude	$\frac{1}{\vert S \vert} \sum_{j \in S} lon_j$
	${}^CF_i^7$	lut_i^{list}	land use type composition	see text
	${}^CF_i^8$	$dlut_i$	dominant land use type	see text
	${}^CF_i^9$	$flut_i^{set}$	first land use types	see text

where $S = \{j \mid j \in \{1, 2, ..., n\} \wedge p^C(V_j) = C_i\}$

Table 5.2.: The features of the supergraph $G^C = (V^C, E^C)$**.** The features of the supernodes V_i^C, representing spatio-temporal clusters of active fire measurements. The land use type composition, lut_i^{list} is a histogram counting the number of occurrences of each land use type within a cluster V_i^C. The dominant land use type, $dlut_i$, is the land use type with the largest number of occurrences within a cluster V_i^C. The first land use types, $flut_i^{set}$, is the set of land use types a given cluster V_i^C has begun on. In most cases - due to the temporal resolution - we can not determine the specific land use type from which a fire cluster started spreading, but only the set of land use types fires occurred on at the first satellite pass.

on all edges, such that $E = \{E_{ij} \mid i, j \in \{1, 2, ..., n\} \wedge E_{ij} \neq \varnothing\}$ leaves only edges between nodes that are neighbours according to the thresholds.

Identifying the connected components of G results in a labelling of the nodes according to their respective cluster membership. A total of $n^C = 168.784$ spatio-temporal clusters were found, and their labels are transferred as features to the nodes of G, $V_i = \{lat_i, lon_i, t_i, lut_i, C_i\}$, where C_i indicates to which cluster a node V_i belongs to. We denote the corresponding partition function by p^C, hence $p^C(V_i) = C_i$. This labelling induces a partition of the graph $G = (V, E)$ into n^C spatio-temporal clusters V_i^C of the supergraph $G^C = (V^C, E^C)$, with $V_i^C = \{V_j \mid j \in \{1, 2, ..., n\} \wedge p^C(V_j) = C_i\}$.

Last, we compute partition-specific features ${}^CF_i^j$ to assign to the supernodes V_i^C, based on the features of the nodes $V_i \in V$. These fire cluster features and their calculation are described in Tab. 5.2. The most important features for the statistical characterization of fire clusters below are:

- **diameter:** the "size" of a cluster, given by the largest great-circle distance between all pairs of active fire measurements within a cluster.
- **lifetime:** the duration of a cluster, given by the time difference between the last and the first fire measurements within a cluster.
- **dominant land use type:** the land use type with the largest number of occurrences within a cluster, given that at least 80% of the measurements belong to that land use type. Otherwise, the dominant land use type "Mixed" is assigned to the cluster.

- **first land use types:** the set of land use types a given cluster has started on. In most cases - due to the temporal resolution - we can not determine the specific land use type from which a fire cluster started spreading, but only the set of land use types fires occurred on at the first satellite pass.

5.6. Statistical Characteristics of Fire Clusters

We first look at the cluster-size distributions of the spatio-temporal fire clusters. The distributions of cluster diameters and lifetimes for the different land use types are depicted in Figs. 5.4(a) and (b), respectively. Clusters with a dominant land use type "Nonforest" clearly dominate in terms of frequency, but also in terms of largest sizes (more than 30 km diameter) and longest lifetimes (more than 100 hours). This is followed by the land use types "Forest", "Pasture", and "Agriculture", which show a consistent decrease in frequency and slopes. Clusters with a dominant land use type "Secondary Vegetation" and "Deforested" are the least frequent, smallest and shortest clusters overall. These observations are consistent with the results of Cano-Crespo et al. (2015), who used a different dataset to determine the burned areas on different land use types. A maximum-likelihood evaluation (analogously to Sec. 4.4.2) reveals that all diameter and lifetime distributions exhibit heavy tails, i.e. none of the distributions' tails are exponentially bounded.

The originating land use types of the different kinds of clusters are investigated next. As mentioned above, we can usually not determine a single land use type a given cluster has started spreading from, due to the temporal resolution. We can, however, determine the set of land use types fires of a given cluster occurred on at the first satellite pass (in the following, referred to as the "first land use types" of a cluster). With this information, we can answer questions such as the following: "How many of all clusters have 'Agriculture' in their first land use types?" or "How many of all clusters with a dominant land use type 'Forest', have 'Pasture' in their first land use types". These questions are of particular interest, since we know that the majority of fires are caused by anthropogenic activities (Cochrane, 2003; Asner et al., 2005; Laurance et al., 2001). Fires are used to convert natural vegetation into pasture and agricultural fields (Morton et al., 2006; Armenteras and Retana, 2012; Davidson et al., 2012), deforested areas are maintained by fires (Cochrane et al., 1999; Barona et al., 2010), and farmers repeatedly burn pasture lands to re-new the grass for their cattle. It has also been shown that fires escaping from managed pastures and agricultural lands significantly contribute to forest fires (Cano-Crespo et al., 2015).

Our findings are summarized in Tab. 5.3. For the interpretation of this table, it is important to know that more than half of all clusters (56.1%) are comprised of one active fire measurement only, and 93.4% are comprised of 5 or less measurements. For this reason, and because of the relatively strict condition on whether a cluster is assigned a dominant land use type (more than 80% of its nodes have to be of the respective land use type), the diagonal values are very large (close to 100% for all land use types), and the off-diagonal values are very small (less than 0.5% for all

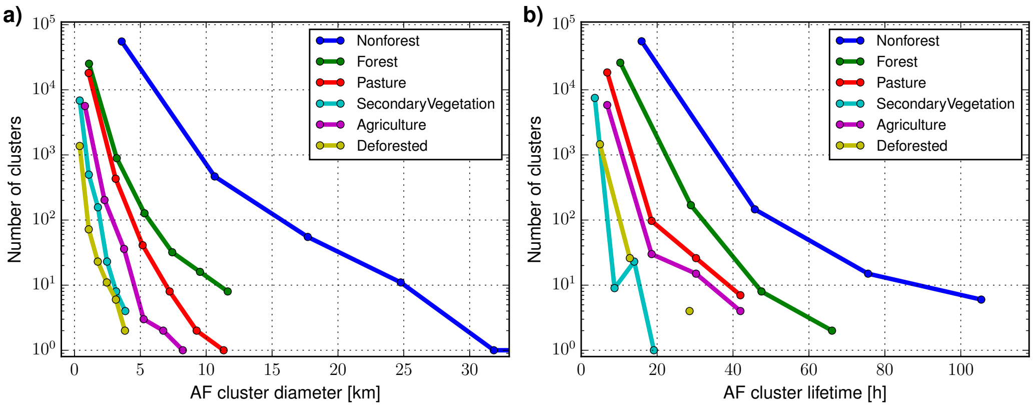


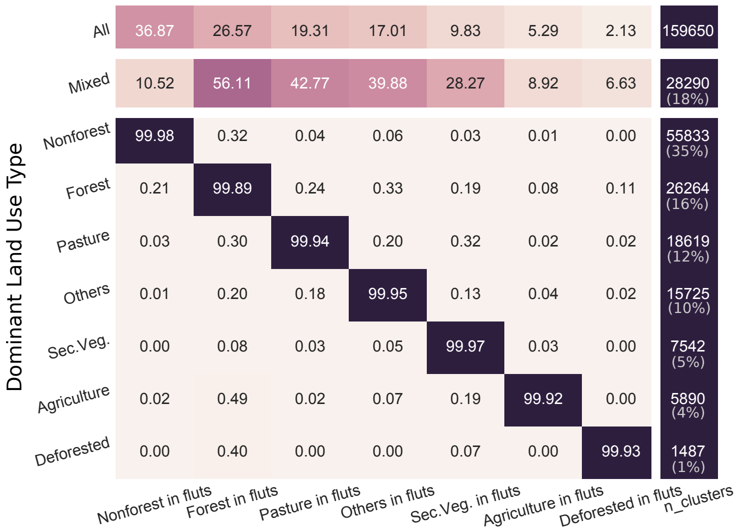
Figure 5.4.: Distributions of cluster diameters and lifetimes for different land use types. **(a)** The frequency distribution of fire cluster diameters (in kilometers) for the different dominant land use types ("Mixed" and "Others" were excluded for visibility). **(b)** The frequency distribution of fire cluster lifetimes (in hours) for the different dominant land use types ("Mixed" and "Others" were excluded for visibility).

combinations). This picture changes noticeably when we restrict the computation of the statistics to clusters with more than 3 or 4 nodes (not shown). Considering all clusters, 19% have "Pasture", 5% have "Agriculture", and 2% have "Deforested" in their first land use types. With regard to the class of "Mixed" clusters, which make up 18% of all clusters, a remarkable 43% of them have "Pasture", 9% have "Agriculture", and 7% have "Deforested" in their first land use types, whereas only 11% have "Nonforest" in their first land use types. This observation asks for a closer inspection of these clusters, in particular with regard to their spatial distribution over the Legal Amazon region. The statistics in Tab. 5.3, however, have to be confirmed with different combinations of distance thresholds, and tested with an appropriate null model.

Finally, we examine the "burning profiles" of the different types of fire clusters. Given the differences between the diameter and lifetime distributions for the different kinds of clusters (see Fig. 5.4), the question arises whether it is possible to predict a cluster's dominant land use type, given solely its diameter (lifetime), or both its diameter and lifetime in conjunction. Considering the lack of land use type data for all years but 2008 and 2010, this would allow us to estimate land use types solely based on the MODIS active fire measurements, which are available from the year 2000 to the present. A byproduct of the probabilistic classifiers derived below are insightful visualizations of the burning characteristics for the different cluster types.

First, we estimate the probability sequence of all of a cluster's possible dominant land use types, given its diameter (lifetime). Using Bayes' Theorem, these conditional probability distributions can be expressed by

$$p(dlut \mid d) = \frac{p(d \mid dlut) \cdot p(dlut)}{p(d)} \tag{5.4}$$



Dominant Land Use Type	Nonforest in fluts	Forest in fluts	Pasture in fluts	Others in fluts	Sec.Veg. in fluts	Agriculture in fluts	Deforested in fluts	n_clusters
All	36.87	26.57	19.31	17.01	9.83	5.29	2.13	159650
Mixed	10.52	56.11	42.77	39.88	28.27	8.92	6.63	28290 (18%)
Nonforest	99.98	0.32	0.04	0.06	0.03	0.01	0.00	55833 (35%)
Forest	0.21	99.89	0.24	0.33	0.19	0.08	0.11	26264 (16%)
Pasture	0.03	0.30	99.94	0.20	0.32	0.02	0.02	18619 (12%)
Others	0.01	0.20	0.18	99.95	0.13	0.04	0.02	15725 (10%)
Sec.Veg.	0.00	0.08	0.03	0.05	99.97	0.03	0.00	7542 (5%)
Agriculture	0.02	0.49	0.02	0.07	0.19	99.92	0.00	5890 (4%)
Deforested	0.00	0.40	0.00	0.00	0.07	0.00	99.93	1487 (1%)

Table 5.3.: Where did fire clusters start spreading? This table answers the following question: given a set of clusters (e.g. all clusters, or only clusters with a dominant land use type "Forest", see rows), how many of them (in percent) started spreading with a certain land use type in their "first land use types (fluts)" set (see columns). The last column contains the number of clusters within each group (and their percentage of the total number of clusters). For instance, given the set of clusters with a dominant land use type "Pasture" (of which there are 18619 in total), the percentage of clusters that had "Forest" in their first land use types is 0.30%.

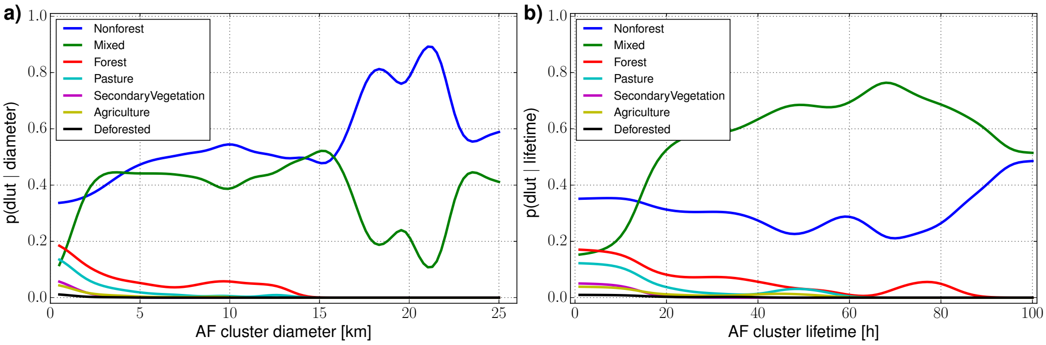


Figure 5.5.: Probability density estimates for dominant land use types given the cluster diameter (lifetime). (a) The probability density estimates for the different land use types given a cluster's diameter, $p(dlut \mid diameter)$, where $dlut$ is one of the dominant land use types. **(b)** The probability density estimates for the different land use types given a cluster's lifetime, $p(dlut \mid lifetime)$, where $dlut$ is one of the dominant land use types.

and

$$p(dlut \mid \Delta t) = \frac{p(\Delta t \mid dlut) \cdot p(dlut)}{p(\Delta t)}, \tag{5.5}$$

where $dlut$ is the dominant land use type, d the diameter, and Δt the lifetime of a cluster. The probability $p(dlut)$ is simply estimated by the proportion of clusters with a dominant land use type $dlut$ (see last column of Tab. 5.3). The probability distributions $p(d)$, $p(d \mid dlut)$ and $p(\Delta t \mid dlut)$ are approximated by a non-parametric gaussian density estimation. The resulting density estimates $p(dlut \mid d)$ and $p(dlut \mid \Delta t)$ are depicted in Figs. 5.5(a) and (b), respectively. For very small diameters (less than one kilometer), the most likely dominant land use type of a cluster is "Nonforest", followed by "Forest", "Pasture" and "Mixed". For increasing diameters, the probability of most land use types rapidly declines, with "Nonforest" and "Mixed" being the exceptions. Only "Forest" sustains a noticeable probability for diameters up to 12-13 kilometers. A similar picture arises for different lifetimes. All land use types except "Nonforest" and "Mixed" decline for increasing lifetimes, albeit less rapidly than for increasing diameters. Noticeably, for lifetimes around 75 hours, we observe a peak in the probability of the "Forest" land use type. Overall, however, given the hypothetical case that we solely know about a cluster's diameter (lifetime), the best guess for any given cluster is consistently "Nonforest" (discarding "Mixed").

Next, we see if the separability of classes is improved by taking into account both a cluster's lifetime and its diameter. Again, using Bayes' Theorem, we can express the probability sequence of all of a cluster's possible dominant land use types by

$$p(dlut \mid d, \Delta t) = \frac{p(d, \Delta t \mid dlut) \cdot p(dlut)}{p(d, \Delta t)}. \tag{5.6}$$

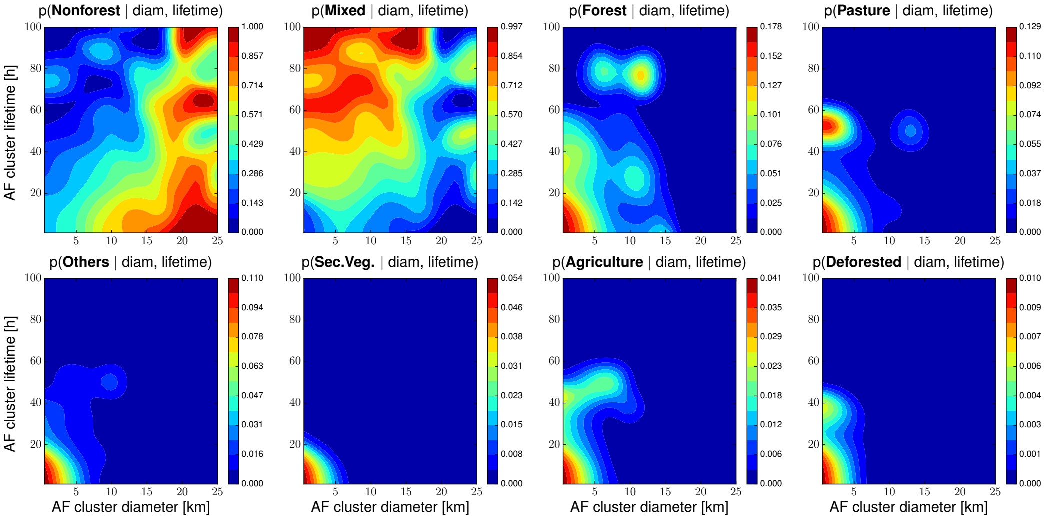


Figure 5.6.: Probability density estimates for dominant land use type given cluster diameter and lifetime. The probability of the different land use types, given a cluster's diameter and lifetime. The x-axis of each subplot represents the diameter (in kilometers), the y-axis represents the lifetime (in hours). The colors indicate the probabilities $p(dlut \mid diam, lifetime)$, where dlut is on of the land use types (see titles). Note that each colorbar is scaled from zero to the respective maximum for a given land use type. This scaling is chosen to improve the visualization of each land use types' burning profile.

The joint probabilities $p(d, \Delta t \mid dlut)$ and $p(d, \Delta t)$ are again approximated by a non-parametric gaussian density estimation. The probability distributions, or "burning profiles", for the different land use types are depicted in Fig. 5.6. Overall - for any given diameter and lifetime combination - the most likely land use types are "Nonforest" or "Mixed". Between these land use types is a split along the diagonal. Large clusters burning for short times are more likely to be "Nonforest" than "Mixed", which means their propagation velocity, on average, exceeds that of "Mixed" clusters. All other cluster types are predominantly small and short-lasting. The "Forest" clusters also exhibit long-lasting (around 80 hours), mid-ranged (5-15 kilometers) instances. Among "Pasture" clusters, there are also long-lasting (45-60 hours), small (up to 5 kilometers) instances. Clusters on "Secondary Vegetation" are exclusively small and short-lasting. Even though we get a more detailed picture of the burning conditions of fire clusters considering diameters and lifetimes in conjunction, the separability of classes has not increased to a point that would improve the probabilistic classifier dramatically. Only if we were to discard the "Mixed" class, two small islands arise in the most-likely land use type matrix (Fig. 5.7) that would not be classified as "Nonforest". In the last section, we discuss several options of how to improve the separability further.

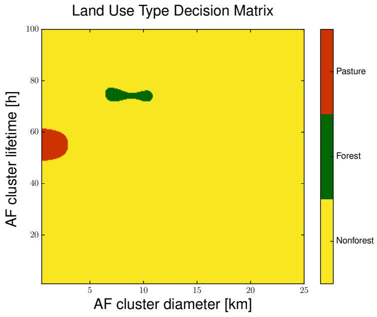


Figure 5.7.: The probabilistic classifier of fire clusters into land use types. Given a cluster's diameter (x-axis) and lifetime (y-axis), the color indicates the most likely dominant land use type according to the conditional probability distribution in Eq. 5.6. The "Mixed" class is discarded in this case.

5.7. Conclusions

In this chapter, we have utilized the deep graph framework to identify spatio-temporal fire clusters in the Legal Amazon region, and then characterized their land use specific burning conditions.

We first combined two high-resolution satellite products - maps of fire-affected areas and land cover maps showing a detailed land use classification - into a graph representation, where nodes correspond to active fire measurements, and edges represent the spatio-temporal proximity between pairs of active fires. After determining reasonable thresholds for the spatial and temporal distances, we identified spatio-temporal fire clusters as the connected components of the graph. For each of the spatio-temporal fire clusters, we calculated a set of features: the diameter, given by the largest great-circle distance between all pairs of active fire measurements within the cluster; the lifetime, given by the time difference between the last and the first fire measurement; the dominant land use type, given by the most frequent land use type within the cluster (given that the land use type makes up at least 80% of the cluster, otherwise the cluster is assigned to a "Mixed" class); and the "first land use types", given by the set of land use types fires of a given cluster occurred on at the first satellite pass.

The distributions of diameters and lifetimes for the different cluster types are dominated by "Nonforest" clusters in terms of frequency, but also largest sizes (more than 30 km diameter) and longest lifetimes (more than 100 hours). This is followed by the land use types "Forest", "Pasture", and "Agriculture", which show a consistent decrease in frequency and slopes. The least frequent, smallest and shortest clusters are "Secondary Vegetation" and "Deforested" clusters. According to a maximum-likelihood evaluation, all distributions exhibit heavy tails, i.e. their tails are not exponentially bounded.

Investigating the originating land use types of the different kinds of clusters, we found that 19% have of all clusters have "Pasture", 5% have "Agriculture", and 2%

have “Deforested” in their first land use types. In the subset of “Mixed” clusters - which make up 18% of all clusters - a remarkable 43% have “Pasture”, 9% have “Agriculture”, and 7% have “Deforested” in their first land use types.

Finally, we derived probabilistic classifiers of fire clusters into dominant land use types, based on either their diameter, their lifetime, or both in conjunction. These classifiers also provided for insightful visualizations of the burning characteristics for the different cluster types. Considering either diameters or lifetimes separately, the probability of most land use types rapidly declines, with “Nonforest” and “Mixed” being the exceptions. “Forest” clusters sustain a noticeable probability for diameters up to 12-13 kilometers, and show a peak for lifetimes around 75 hours. Overall, given the case we would solely know about a cluster’s diameter (lifetime), the best guess for any given cluster is consistently “Nonforest” (discarding “Mixed”). We then tried to increase the separability of classes by taking into account both a cluster’s lifetime and its diameter. We found that large clusters burning for short times are more likely to be “Nonforest” than “Mixed”, which means their propagation velocity, on average, exceeds that of “Mixed” clusters. The other cluster types are predominantly small and short-lasting, while “Forest” clusters also exhibit long-lasting (around 80 hours), mid-ranged (5-15 kilometers) instances, and “Pasture” clusters also show long-lasting (45-60 hours), small (up to 5 kilometers) instances. Even though the separability of cluster classes has not improved dramatically, the decision matrix (Fig. 5.7) shows two small islands. One where “Pasture” is more likely than “Nonforest” and one where “Forest” is more likely than “Nonforest”.

5.8. Outlook

Although our graph-based approach to study the land use specific characteristics of fire clusters is novel and preliminary results seem promising, some work is still left to be conducted, and a few ideas are worth incorporating into the analysis in the future.

Most importantly, we need to fine-tune the spatio-temporal distance thresholds. This could be accomplished, for instance, by comparing the diameter and lifetime distributions with other datasets, such as the burned-area distributions analyzed by Cano-Crespo et al. (2015).

The observation that a large proportion of “Mixed” clusters originate from “Pasture”, “Agriculture” and “Deforested” fields supports the findings of Cano-Crespo et al. (2015), stating that fires from these fields often escape and burn further areas. Studying these particular clusters with respect to their temporal and spatial distribution over the Legal Amazon area could lead to further insights.

With regard to the probabilistic classifier, a number of considerations to improve its predictive power come to mind. We could, for instance, look for differences in the burning profiles of the different cluster types in the dry and wet seasons. The recurrence rate of fires on a given spatial location could help us identify pasture fields, since farmers often burn their fields repeatedly to re-new the grass for their cattle. A very promising extension of the analysis would be to incorporate high-resolution lightning data (Lay et al., 2007), match it with the fire clusters and thereby improve our estimation whether a fire was caused by anthropogenic activities or natural causes.

Chapter 6.

Synchronizability in Noisy Complex Networks

6.1. Summary

The effects of white noise and global coupling strength on the maximum degree of synchronization in complex networks are explored. We perform numerical simulations of generic oscillator models with both linear and non-linear coupling functions on a broad spectrum of network topologies. The oscillator models include the Fitzhugh-Nagumo model, the Izhikevich model and the Kuramoto phase oscillator model. The network topologies range from regular, random and highly modular networks to scale-free and small-world networks, with both directed and undirected edges. We then study the dependency of the maximum degree of synchronization on the global coupling strength and the noise intensity. We find a general scaling of the synchronizability, and quantify its validity by fitting a regression model to the numerical data. This chapter is based on the associated publication P1, and the following sections will closely follow this publication.

6.2. Introduction

The emergence of collective and synchronous dynamics in large ensembles of coupled units is an ubiquitous phenomenon in nature and engineering. Its study has attracted much attention in a variety of fields, such as neuroscience, biology, physics, chemistry or social sciences (Kuramoto, 1984; Pikovsky et al., 2003). For instance, there are proliferating indications that strong synchronization on large scales is related to pathological conditions of the human brain, e.g., epileptic seizures and Parkinson disease (Stam, 2005). Subject of current research includes the comprehension of common properties of network synchronization in dependence on the individual node dynamics, the network topology, internode coupling types and the influence of various types of noise (Barahona and Pecora, 2002; Nishikawa et al., 2003; Motter et al., 2005; Bag et al., 2007). In the context of physiological networks and network medicine, for instance, a strong relationship between the topology of physiological networks and their physiological functions has recently been observed (Bashan et al., 2012; Ivanov and Bartsch, 2014; Bartsch and Ivanov, 2014).

The focus of this study is the influence of uncorrelated white noise on the maximum degree of synchronization. While many studies investigated the phase transition associated with the onset of macroscopic synchronization (Sakaguchi, 1988; Sonnenschein and Schimansky-Geier, 2013), here we are particularly interested in the synchronizability in the aftermath of the phase transition. The effect of noise on the phase-synchronization of non-linear oscillators has for example been studied by Xu et al. (2006), and it is known that white noise prohibits the capability of a system to achieve full synchronization by decreasing the maximum degree of synchronization (Bag et al., 2007). It is not clear, however, how this decrease of the synchronizability relates to the noise intensity. Furthermore, it has been shown that the network topology has a great influence on the time-evolution of local patterns of synchronization on the path towards global coherence (Gómez-Gardeñes et al., 2007). The question, though, whether the topology of the network has an influence on the maximum degree of synchronization, has not been answered yet.

In order to answer these questions, we develop a numerical simulation framework and study the dependency of the maximum degree of synchronization on the global coupling strength and the noise intensity. The framework incorporates three basic types of well known oscillators, namely the Fitzhugh-Nagumo model, the Izhikevich model and the Kuramoto phase oscillator. The oscillators are coupled by both linear and non-linear coupling functions. The coupling topologies include regular, random, small-world, scale-free and highly modular networks, with both directed and undirected edges.

We find a general scaling of the maximum degree of synchronization, and quantify its validity by fitting a regression model to the numerical data.

6.3. The Models

In the following, we will introduce three different oscillator models used in our simulations. The models are widely known and were chosen such that we obtain a diverse set of oscillator models. The Kuramoto model is an ubiquitous phase oscillator model, simple enough to be mathematically tractable, yet sufficiently complex to display a large diversity of synchronization patterns. It is flexible enough to be adapted to many different contexts (e.g. biological models, associative memory models and laser arrays (Acebrón et al., 2005)). The Fitzhugh-Nagumo model provides a simple yet basic representation of firing dynamics and has been broadly used as a model for cardiac cells and spiking neurons (Koch, 1999; Glass et al., 1991). The Izhikevich model is a biologically plausible neuron model, capable of reproducing spiking and bursting behaviour of known types of neocortical and thalamic neurons (Izhikevich, 2003). The parameters of the Izhikevich model were chosen such that it reproduces a bursting behaviour, in order to set its dynamics further apart from the Fitzhugh-Nagumo model and thus expanding the diversity of the oscillator models included in this study.

6.3.1. Kuramoto Model

The stochastic Kuramoto model (Kuramoto, 1984) for N coupled and identical phase oscillators ($i = 1, ..., N$) is described by:

$$\frac{d}{dt}\theta_i(t) = \omega + \xi_i + \frac{g}{\langle k \rangle} \sum_{i=1}^{N} M_{ji} \sin(\theta_j - \theta_i) \tag{6.1}$$

where θ_i is the phase of the i-th oscillator, $\omega = 2\pi$ is its associated natural frequency, g the global coupling strength, $\langle k \rangle$ the average graph connectivity ($\langle k \rangle \equiv \frac{2L}{N}$, with L denoting the total number of (weighted) links). ξ_i stands for Gaussian uncorrelated white noise sources with expectation

$$\mathbb{E}(\xi_i^{in}(t)) = 0 \tag{6.2}$$

and covariance

$$cov(\xi_i^{in}(s), \xi_j^{in}(t)) = 2D^{in}\delta_{ij}\delta(s-t) \ , \tag{6.3}$$

where D^{in} will be referred to as the noise level. M_{ij} is the (weighted and/or directed) adjacency matrix of the simulated network. A weighted link is a link associated with a scalar value, quantifying properties as for instance the frequency of contact between actors in social networks, or the number of synapses connecting a pair of neurons in neural networks. Additionally, one distinguishes between directed and undirected graphs, that is, by whether the edges possess directional information or not, respectively.

6.3.2. Izhikevich Model

The Izhikevich model (Izhikevich, 2003) for N coupled oscillators ($i = 1, ..., N$) is described by the ordinary differential equations:

$$\frac{d}{dt}v_i(t) = 0.04v_i^2 + 5v_i + 140 - u_i + I_0 + I_i \tag{6.4}$$

$$\frac{d}{dt}u_i(t) = a(bv_i - u_i) + \xi_i^{in} \tag{6.5}$$

with an after-spike resetting:

$$\text{if } v_i \geq 30, \text{ then } v_i \text{ is set to } c \text{ and } u_i \text{ is set to } u_i + d \tag{6.6}$$

According to Izhikevich (2004), v_i represents the membrane potential and u_i a membrane recovery variable. $\xi_i^{in}(t)$ are white noise sources as in Eqs. (6.2) and (6.3), where the noise level D^{in} represents the intrinsic noise of an isolated neuron (e.g. ionic conductance noise, ionic pump noise). Synaptic currents are delivered via the variable I_i, stated below under coupling functions. After a spike reaches its maximum ($v = +30$), both variables are reset according to Equation (6.6). By assigning different

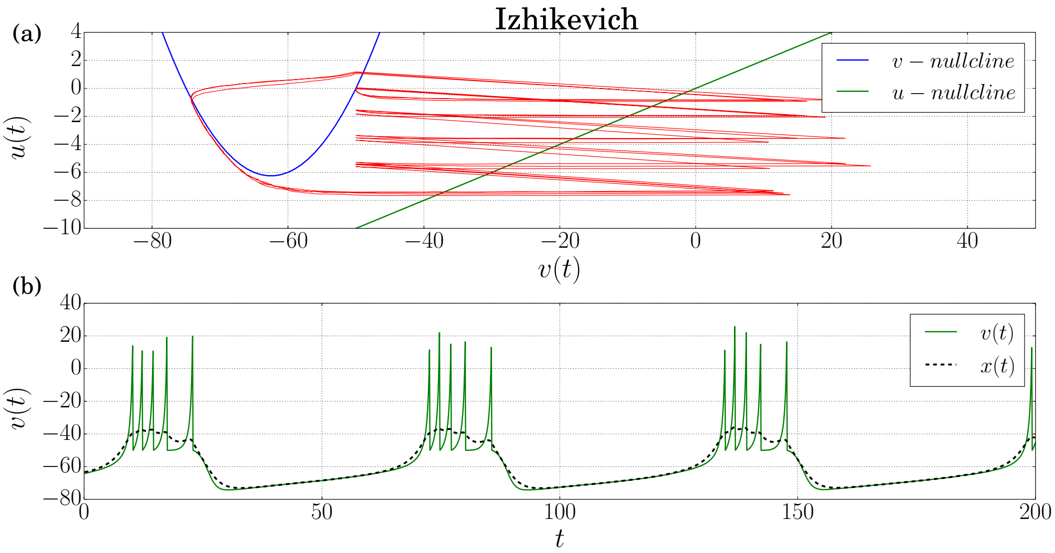


Figure 6.1.: Phase space portrait of a single Izhikevich neuron and corresponding time-series and low-pass filtered time-series of the membrane potential v(t)**. (a)** The two dimensional phase space trajectory of an uncoupled Izhikevich neuron, subject to Gaussian noise as given in Eqs. (6.2) and (6.3) with a noise intensity of $D^{in} = 0.05$ is drawn in red. The two nullclines are drawn in green and blue. **(b)** The corresponding time-series of the membrane potential $v(t)$ of the uncoupled Izhikevich neuron is drawn in green, the dashed black line represents the low-pass filtered time-series $x(t)$ of the membrane potential (see Equation (6.15)).

values to the parameters (a, b, c, d), the model can reproduce spiking and bursting behaviour of various known types of neocortical and thalamic neurons (Izhikevich, 2003). The parameters $a = 0.02, b = 0.2, c = -50.0, d = 2.0$ and $I_0 = 10.0$ are chosen such that the model generates a bursting signal, as depicted in Fig. 6.1.

6.3.3. Fitzhugh-Nagumo Model

The second neurological model we consider is the Fitzhugh-Nagumo model (Fitzhugh, 1961) for N coupled oscillators ($i = 1, ..., N$). It is composed of the following differential equations:

$$\frac{d}{dt}v_i(t) = v_i - \frac{v_i^3}{3} - u_i + I_0 + I_i + \xi_i^{in} \tag{6.7}$$

$$\frac{d}{dt}u_i(t) = \frac{v_i - a - bu_i}{\tau} \tag{6.8}$$

The variable v_i represents the membrane potential and u_i the recovery variable for the neuron membrane potential. Again, synaptic currents are transmitted via I_i and ξ_i^{in} stands for Gaussian white noise sources as in Eqs. (6.2) and (6.3). The constants

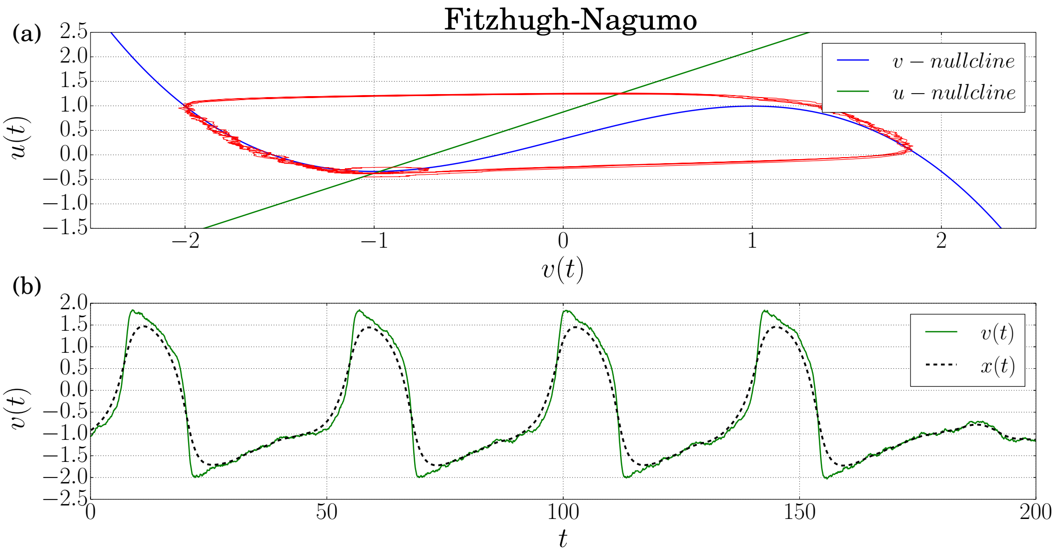


Figure 6.2.: Phase space portrait of a single Fitzhugh-Nagumo neuron and corresponding time-series and low-pass filtered time-series of the membrane potential v(t). **(a)** The two dimensional phase space trajectory of an uncoupled Fitzhugh-Nagumo neuron, subject to Gaussian noise as given in Eqs. (6.2) and (6.3) with a noise intensity of $D^{in} = 0.05$ is drawn in red. The two nullclines are drawn in green and blue. **(b)** The corresponding time-series of the membrane potential $v(t)$ of the uncoupled Fitzhugh-Nagumo neuron is drawn in green, the dashed black line represents the low-pass filtered time-series $x(t)$ of the membrane potential (see Eq. (6.15)).

$a = -0.7, b = 0.8, \tau = 12.5$ and $I_0 = 0.328$ are chosen such that the neuron is spiking continuously, as illustrated in Fig. 6.2.

6.3.4. Coupling Functions

For the Izhikevich and the Fitzhugh-Nagumo model, simulations are performed with both electrical and chemical coupling functions. Hence, I_i in Eqs. (6.4) and (6.7) takes the form:

$$I_i = \xi_i^{ex} + I_{el,i} + I_{chem,i} \tag{6.9}$$

where ξ_i^{ex} stands again for Gaussian uncorrelated white noise sources with expectation

$$\mathbb{E}(\xi_i^{ex}(t)) = 0 \tag{6.10}$$

and covariance

$$cov(\xi_i^{ex}(s), \xi_j^{ex}(t)) = 2D^{ex}\delta_{ij}\delta(s-t) \ , \tag{6.11}$$

but here it is accounting for extrinsic noise sources like gap junctions and chemical synapses (e.g. synaptic release noise).

Electrical Coupling

The electrical transmission, $I_{el,i}$ in Eq. (6.9) can be realized with a linear function of the form (Pinto et al., 2000):

$$I_{el,i}(v_i, \mathbf{v}) = \frac{g_{el}}{N} \sum_{j=1}^{N} M_{ji}(v_j - v_i) \tag{6.12}$$

where g_{el} represents the global electrical coupling strength, v_i and v_j stand for the membrane potentials of the post-synaptic and the pre-synaptic neurons respectively, and N is the number of neurons in the network. The local coupling strength between two connected neurons is obtained by the weighted adjacency matrix $\mathbf{M}$ of the simulated network.

Chemical Coupling

There are several ways of modelling chemical synapses. Pinto et al. (2000), for example, use the approach of adding a first order dynamic for each synapse. A computational more effective implementation which still conserves the crucial properties is given by the following function (Belykh et al., 2005):

$$I_{chem,i}(v_i, \mathbf{v}) = \frac{g_{ch}}{N}(V_s - v_i) \sum_{j}^{N} M_{ji}\Gamma(v_j - \Theta) \tag{6.13}$$

with

$$\Gamma(x) = (1 + \exp[-\lambda x])^{-1} \tag{6.14}$$

where g_{ch} is the global coupling strength for chemical synapses. The local coupling strengths M_{ij}, N, v_i and v_j are as described for the electrical coupling. The sigmoidal function Γ represents the thresholding behaviour of the synapse, with λ being the control parameter of the steepness. The multiplier $(V_s - v_i)$ reduces the input, if the post-synaptic neuron itself is already depolarized. The parameters for the Izhikevich model are $V_s = 30.0$, $\lambda = 0.41$ and $\Theta = -50.0$, and for the Fitzhugh-Nagumo model $V_s = 1.75$, $\lambda = 11.58$ and $\Theta = -0.5$ respectively. Although taking into account several aspects of chemical synapses, some properties like transmission-delay are neglected in this model.

6.4. The Networks

We have chosen six different networks from a broad spectrum of topologies, ranging from all-to-all connectivity over regular, scale-free, small-world and modular networks to a random topology, including directed and undirected edges. We did so in order to thoroughly investigate the influence of the network topology on the maximal degree of synchronization and to test the generality of the scaling as introduced below.

a) An unweighted and undirected all-to-all network **A**, consisting of 256 nodes where every node is connected with any other node (global-coupling topology), as depicted in Fig. 6.3(a).

b) An Erdös-Rény random graph (Erdős and Rényi, 1959) **R**, consisting of 256 nodes connected by 1015 undirected and unweighted edges, illustrated in Fig. 6.3(b).

c) Figure 6.3(c) shows a computer-generated graph with two hierarchical levels of communities, as proposed by Arenas et al. (2006). The undirected and unweighted network **H** consists of 256 nodes and is structured into two predefined hierarchical community levels. The inner communities consist of 16 nodes each and the outer communities consist of 64 nodes each. Each node has 13 links within its inner community, four links within its outer community and one more link with any other randomly chosen node in the network, adding up to a total of 1015 links in the entire network.

d) Furthermore, we have selected the real-world network of the somatic nervous system of the soil nematode *C.elegans*. The nervous system of *C.elegans* is the only one that has been almost completely mapped down to the synaptic level, and shares properties of small-world and scale-free networks (Varshney et al., 2011). The data is based on the most complete database to date, provided by Varshney et al. (2011). It is composed of two adjacency matrices. The electrical synapse network **G**, undirected and weighted, connecting the 279 somatic neurons by a total of 887 (514, discarding weights) gap junctions and the chemical synapse network **S**, directed and weighted, connecting the neurons by a total of 6394 (2194, discarding weights) chemical synapses. The weight of a pair of connected neurons reflects the number of electrical (chemical) synapses connecting it. The simulations are performed on the combined network $\mathbf{C} = \mathbf{G} + \mathbf{S}$, depicted in Fig. 6.3(d) (weights have been discarded in the Figure), which is simply the sum of the two adjacency matrices of the gap junction and the chemical synapse network. Gap junctions are thus treated as double-sided directed connections. This combined network consists of 279 nodes and 8168 (2990, discarding weights) directed connections.

e) The unweighted realization of this network, as depicted in Fig. 6.3(e), will be referred to as **U**.

f) Furthermore, simulations are performed on a rewired surrogate network of the unweighted graph of *C.elegans* **U**, where only the degree-distribution is preserved. It is obtained by iteratively swapping randomly selected edges (Rao et al., 1996) of **U**. At each iteration, two links are chosen at random $((n1 \mapsto n2)$ and $(n3 \mapsto n4))$ and rewired $((n1 \mapsto n4)$ and $(n3 \mapsto n2))$, unless the respective new links do not already exist or introduce self-loops. Repeating this process often enough, all internal structure of the

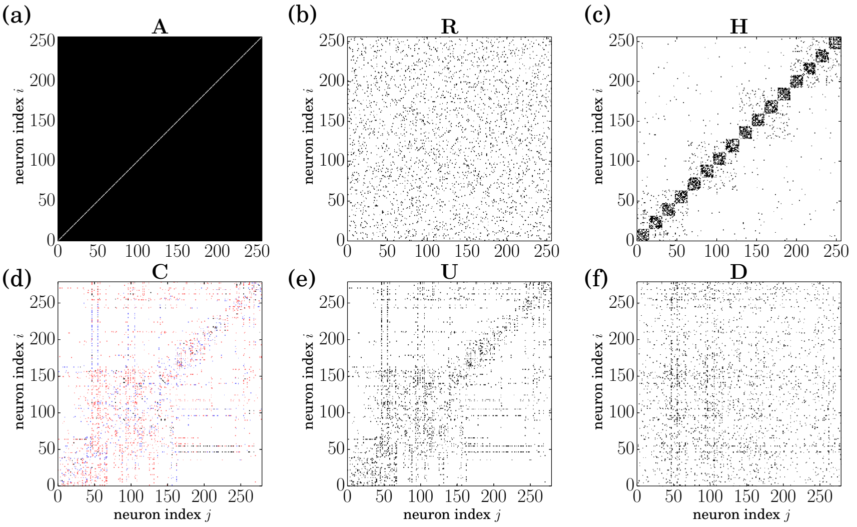


Figure 6.3.: Adjacency matrices of simulated networks. (a) Adjacency matrix of the all-to-all global-coupling network $\mathbf{A}$. **(b)** Adjacency matrix of the Erdös-Rény random graph $\mathbf{R}$, consisting of 256 nodes connected by 1015 undirected and unweighted edges. **(c)** Adjacency matrix $\mathbf{H}$ of the synthetically structured modular network. **(d)** Adjacency matrix $\mathbf{C}$ of the directed and weighted (weights are not shown in the Figure) real-world network of *C.elegans*, where gap junctions are coloured blue, chemical synapses are coloured red and coinciding gap junctions and chemical synapses are coloured black, respectively. **(e)** Adjacency matrix of the directed and unweighted realization of *C.elegans'* network $\mathbf{U}$. **(f)** Adjacency matrix $\mathbf{D}$ of the degree-matched surrogate network of *C.elegans*.

original network is destroyed, except the degree-distribution. The random, directed and unweighted network, degree-matched to the unweighted combined network of *C.elegans* $\mathbf{U}$ will be referred to as $\mathbf{D}$. Its adjacency matrix is shown in Fig. 6.3(f).

6.5. Numerical Simulation Setup

In order to quantify the decrease in maximum synchronization of complex networks of oscillators due to Gaussian white noise and in dependence on the global coupling strength, we develop a simulation framework, as described in the following.

For the Izhikevich and the Fitzhugh-Nagumo model, the stochastic differential equations are solved by a standard first-step Euler method. Sufficient accuracy was achieved with a step size of $\Delta t = 0.1$. Note that, in all simulations, the intrinsic noise D^{in} and the extrinsic noise D^{ex} were set to equal values, therefore the noise level will simply be referred to as $D \equiv D^{in} = D^{ex}$. To avoid a priori synchronizations, initial conditions are drawn randomly from a uniform distribution and 50.000 iterations are calculated.

For the Izhikevich model, the generated time-series mainly contain two frequencies, a fast occurrence of spikes and a slow occurrence of bursts, as can be seen in Fig. 6.1(b). Since we are only interested in the synchronization of bursting activity, the following low-pass filter was applied to the signal:

$$x_{i,t=0} := v_{i,t=0} \tag{6.15}$$

$$x_{i,t} = av_{i,t} + (1-a)x_{i,t-dt} \tag{6.16}$$

with $a = 0.90$. The time-series generated with the Fitzhugh-Nagumo model were smoothed by the low pass filter as well. To measure the synchronicity between two time-series, x_i and x_j, the Pearson correlation coefficient is calculated for the low pass filtered signal of each pair of neurons in the network. It is defined as:

$$R_{ij} = \frac{\sum_t (x_{i,t} - \mu(x_i))(x_{j,t} - \mu(x_j))}{\sigma(x_i)\sigma(x_j)} \tag{6.17}$$

where $\mu(x)$ is the mean value and $\sigma(x)$ the standard deviation of the time-series.

Because the initial conditions are chosen randomly, they are not necessarily close to the system's attractor. The transient is therefore discarded, and the Pearson correlation coefficient is calculated from the $20.000th$ iteration onwards. Furthermore, to eliminate random synchronizations, for each set of parameters (network, model, coupling method, global coupling strength and noise level) ten realizations are calculated and the arithmetic mean of all realizations $\overline{R}_{ij} = \frac{1}{10}\sum_{r=1}^{10} R_{ij}^r$ is taken. To quantify the average synchronization of the entire network, the absolute mean correlation $\langle\overline{\mathbf{R}}\rangle = \frac{1}{N(N-1)} \sum\limits_{i,j,i\neq j}^{N} |\bar{R}_{ij}|$ is calculated.

For the Kuramoto model, a standard first-step Euler method is implemented as well, but sufficient accuracy was only established with a step size of $\Delta t = 0.01$. Initial conditions are drawn from a uniform distribution in the interval $[0, 2\pi]$, and 200.000 iterations are calculated. To measure the level of synchronization for system (6.1), we take the classical order parameter $r(t) = \frac{1}{N}|\sum_{j=1}^{N} e^{i\theta_j(t)}|$ and average it over the last 30.000 iterations, $O = \langle r(t)\rangle_T$. Again, for each set of parameters (network, global coupling strength and noise level) we integrate ten realizations and take the arithmetic mean, $\overline{O} = \frac{1}{10}\sum_{r=1}^{10} O^r$.

6.6. Results

We are interested in the influence of the noise level D and the global coupling strength g on the mean correlation $\langle\overline{\mathbf{R}}\rangle$ for the Izhikevich and the Fitzhugh-Nagumo model, and the order parameter $\overline{O}$ for the Kuramoto model, respectively. For the sake of convenience, the order parameter for a simulation with the Kuramoto model, $\overline{O}$, will be referred to as $\langle\overline{\mathbf{R}}\rangle$ as well, $\overline{O} \equiv \langle\overline{\mathbf{R}}\rangle$. The following analysis of the simulations is based on the interpretation of the mean correlation as a function of the noise level and the global coupling strength: $\langle\overline{\mathbf{R}}\rangle = \langle\overline{\mathbf{R}}\rangle\,(g, D)$

Model	Network	Coupling Method	Nr. of Simulations	NRMSD (linear)	NRMSD
Izhikevich	**H**	electrical	261	8.0%	5.2%
		chemical	195	9.0%	5.8%
	C	electrical	300	10.6%	6.4%
		chemical	360	11.9%	7.9%
Fitzhugh-Nagumo	**H**	electrical	171	7.9%	5.8%
		chemical	171	7.7%	5.7%
	C	electrical	153	7.5%	3.8%
		chemical	81	7.7%	3.9%
	R	electrical	390	6.7%	4.9%
	A	electrical	320	7.5%	5.0%
	U	electrical	126	4.4%	3.2%
	D	electrical	126	4.4%	4.0%
Kuramoto	**H**	sinusoidal	70	7.7%	5.5%
	C	sinusoidal	112	6.6%	2.5%
	R	sinusoidal	70	7.3%	4.7%
	A	sinusoidal	70	7.8%	4.6%
	D	sinusoidal	88	6.2%	3.5%
			3064	7.6% ± 1.8%	4.8% ± 1.3%

Table 6.1.: Division of all numerical simulations. The first column states the model of the simulations, followed by the underlying network, the coupling method and the number of simulations. The fifth column quotes the normalized root-mean-square deviation (NRMSD) of a fit to the linear model $R^{\star,lin}_{(g,D,\mathbf{w}^l)}$ as described in Eq. (6.20), and the last column quotes the NRMSD of a fit to the non-linear model $R^{\star,nonlin}_{(g,D,\mathbf{w}^{nl})}$ as described in Eq. (6.21). In the bottom row, the total number of simulations is stated, followed by the mean NRMSD over all simulation setups (± one standard deviation) for the linear model and the non-linear model.

In total, 3064 simulations were performed, each corresponding to a specific model, coupling method, network, global coupling strength and noise level. They are divided into 17 subsets of data as shown in Tab. 6.1.

In Figs. 6.4(a) and 6.4(b), the dependency of the average synchronization $\langle\overline{\mathbf{R}}\rangle$ on the coupling strength g and the noise level D is shown for the all-to-all network **A** of electrically coupled Fitzhugh-Nagumo neurons and for the degree-matched random network of *C.elegans*, **D**, of Kuramoto phase oscillators, respectively. The result is a two-dimensional surface in three-dimensional euclidean space, where every grid point on the surface represents the mean of ten realizations of a simulation and is associated with the corresponding mean correlation $\langle\overline{\mathbf{R}}\rangle$. Lines of constant noise level D are projected on the $(g\langle\overline{\mathbf{R}}\rangle)$-plane (green lines), lines of constant coupling strength g are projected on the $(D\langle\overline{\mathbf{R}}\rangle)$-plane (red lines) and lines of constant mean correlation $\langle\overline{\mathbf{R}}\rangle$ are projected on the (gD)-plane (blue lines).

Comparing the shape of the surfaces described by the correlation function $\langle\overline{\mathbf{R}}\rangle = \langle\overline{\mathbf{R}}\rangle\,(g, D)$, it becomes apparent that they closely resemble each other (note that the input data, the global coupling strength g and the noise level D were rescaled to

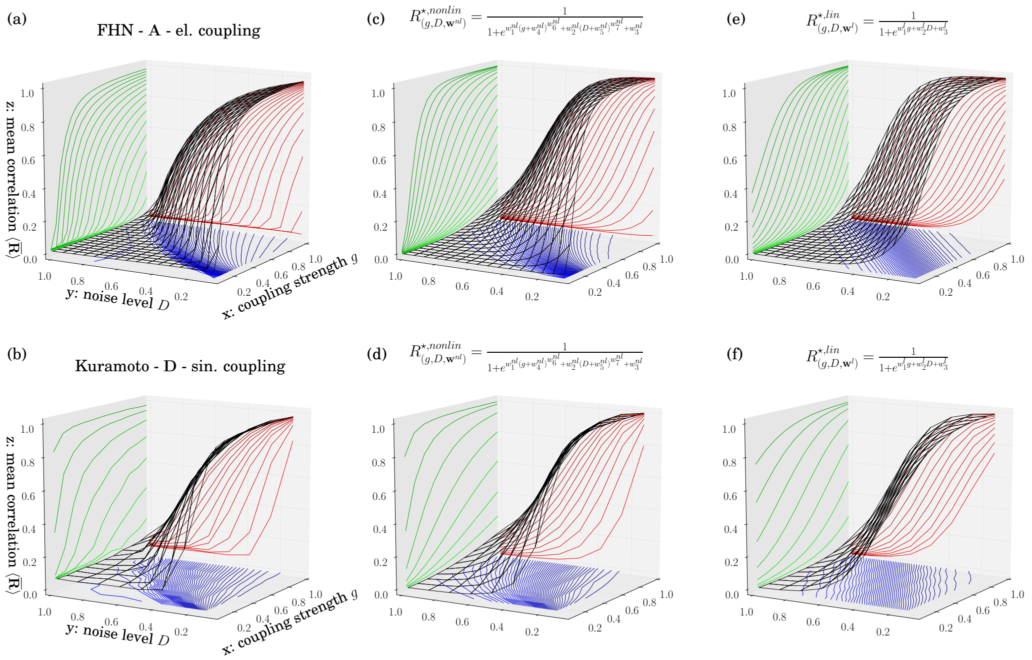


Figure 6.4.: Average synchronization $\langle\overline{\mathbf{R}}\rangle$ in dependence of the noise level D and the coupling strength g_{el} for the numerical simulations (a)-(b) and the regression models (c)-(f). (a)-(b) The results of numerical simulations for the network $\mathbf{A}$ of electrically coupled Fitzhugh-Nagumo neurons and the network $\mathbf{D}$ of Kuramoto phase oscillators respectively. Lines of constant noise level D are drawn in green, lines of constant coupling strength g are drawn in red and lines of constant mean correlation $\langle\overline{\mathbf{R}}\rangle$ are drawn in blue. **(c)-(d)** Fit to the numerical data with the non-linear regression model $R^{\star,nonlin}_{(g,D,\mathbf{w}^{nl})}$ as described in Eq. (6.21), for the Fitzhugh-Nagumo model and the Kuramoto model respectively. **(e)-(f)** Fit to the numerical data with the linear regression model $R^{\star,lin}_{(g,D,\mathbf{w}^{l})}$ as described in Eq. (6.20), for the Fitzhugh-Nagumo model and the Kuramoto model respectively.

the range $]0, 1]$). For constant noise levels D the mean correlation $\langle\overline{\mathbf{R}}\rangle$ resembles a sigmoid function, where the steepness of these curves declines with an increasing noise level, and the inflection point moves towards larger values of g. The lines of constant coupling strength (red lines) resemble sigmoid curves as well, and for increasing values of the coupling strength, the steepness of the sigmoids declines, and the inflection point moves towards higher values of D. The blue lines, representing intersection lines of different (gD)-planes with the surface described by the correlation function, show how the coupling strength scales with the noise for fixed values of the mean correlation. The scaling seems to be of the form $g \sim D^{\beta}$, with $\beta \approx 1$ or slightly larger.

Remarkably, the same behaviour is observed independently of the models, coupling methods and networks numerical simulations were conducted with (not all Figures shown). In order to quantify the deviation of the numerical data from the observed functional dependency, we now introduce two regression models, which will be fit to the numerical data by a least squares method, and the normalized root-mean-square deviation (NRMSD) will serve as a measure for the difference between the observed values of $\langle\overline{\mathbf{R}}\rangle$ and the values implied by the regression models. The NRMSD is defined as the square root of the mean square error (MSE) normalized by the range of observed values:

$$\mathrm{MSE} = \frac{1}{p}\sum_{n=1}^{p}(R^{\star}_{(g^{(n)},D^{(n)},\mathbf{w})} - \langle\overline{\mathbf{R}}\rangle^{(n)})^2 \tag{6.18}$$

$$\mathrm{NRMSD} = \frac{\sqrt{\mathrm{MSE}}}{\langle\overline{\mathbf{R}}\rangle_{max} - \langle\overline{\mathbf{R}}\rangle_{min}} \tag{6.19}$$

where the sum in Eq. (6.18) is taken over all data points of a simulation setup, and $R^{\star}_{(g^{(n)},D^{(n)},\mathbf{w})}$ is the model output for a given $g^{(n)}$ and $D^{(n)}$. The optimal model parameters are given by the vector $\mathbf{w}$, which is retrieved by a gradient descent method.

The first regression model is of rather low complexity, based only on the observation that the dependency of the coupling strength on the noise level for fixed values of $\langle\overline{\mathbf{R}}\rangle$ is close to linear, and the sigmoidal dependencies of $\langle\overline{\mathbf{R}}\rangle$ on g (D) for fixed values of D (g). It is a two-dimensional sigmoid function, given by:

$$R^{\star,lin}_{(g,D,\mathbf{w}^l)} = \frac{1}{1+e^{w_1^l g + w_2^l D + w_3^l}} \tag{6.20}$$

where $\mathbf{w}^l = (w_1^l, w_2^l, w_3^l)$ are the model parameters, which were fit separately for each simulation setup (given by the oscillator model, coupling method and network topology).

The second regression model is of much higher complexity, and supposedly has the capability of fitting the data very well. It serves to give a lower bound on the NRMSD, in order to classify to goodness of the fit of the linear regression model. It is given by the function:

$$R^{\star,nonlin}_{(g,D,\mathbf{w}^{nl})} = \frac{1}{1 + e^{w_1^{nl}(g+w_4^{nl})^{w_6^{nl}} + w_2^{nl}(D+w_5^{nl})^{w_7^{nl}} + w_3^{nl}}} \tag{6.21}$$

where $\mathbf{w}^{nl} = (w_1^{nl}, w_2^{nl}, w_3^{nl}, w_4^{nl}, w_5^{nl}, w_6^{nl})$ are the model parameters, which again were fit separately for each simulation setup.

In Tab. 6.1, the NRMSDs for both regression models and all simulation setups are summarized. The average NRMSD of all setups for the non-linear regression model has, as expected, a very low value of $\text{NRMSD}^{nonlin} = 4.8\% \pm 1.3\%$ and is therefore capable of fitting the data very well. With a value of $\text{NRMSD}^{lin} = 7.6\% \pm 1.8\%$, the average NRMSD for the linear model is about 1.6 times larger, but considering the simplicity of the model, its capability of fitting the data is surprisingly good.

6.7. Discussion and Conclusion

In summary, we performed numerical simulations on a variety of oscillator models in a broad spectrum of network topologies, coupled by both linear and non-linear coupling functions. The models range from the Fitzhugh-Nagumo model in a continuously spiking state, over the Izhikevich model in a periodically bursting state to the Kuramoto model of identical phase oscillators. The network topologies include scale-free, small-world, regular, random and highly modular networks, with both directed and undirected edges.

We were interested in the maximum degree of synchronization, in dependence on the global coupling strength g and the intensity of the white noise sources D. We found common characteristics independent of the oscillator model, network and coupling type. For fixed values of the noise intensity, we found a sigmoidal dependency of the synchronizability on the global coupling strength, and for fixed values of the global coupling strength, the dependency of the synchronizability on the noise intensity seems to be sigmoidal too. Furthermore, the scaling of the noise intensity with the global coupling strength for fixed values of the average synchronization of a network seems to be close to linear.

Introducing a regression model of the form $R^{\star,lin}_{(g,D,\mathbf{w}^l)} = (1+e^{w_1^l g + w_2^l D + w_3^l})^{-1}$ allowed us to quantify the deviation between the numerical results and the proposed sigmoidal dependency for each simulation setup in terms of the normalized root-mean-square deviation (NRMSD), as stated in Eq.(6.19). Given the simplicity of the regression model, and the diversity of oscillator models, networks and coupling methods, the consistently small NRMSDs ($\text{NRMSD}^{lin} = 7.6\% \pm 1.8\%$, see Tab. 6.1) throughout all setups are rather unexpected.

The considerably more complex regression model given by $R^{\star,nonlin}_{(g,D,\mathbf{w}^{nl})} = (1 + e^{w_1^{nl}(g+w_4^{nl})^{w_6^{nl}} + w_2^{nl}(D+w_5^{nl})^{w_7^{nl}} + w_3^{nl}})^{-1}$, capable of reproducing a non-linear relationship between noise intensity and coupling strength for fixed values of the mean correlation as well as a rising slope of that relation for increasing values of the mean correlation, only diminishes the average NRMSD to $\text{NRMSD}^{nonlin} = 4.8\% \pm 1.3\%$.

We have thus shown that the maximum degree of synchronization can be approximated quite adequately by a simple 2-dimensional sigmoidal function with a linear relation between noise intensity and global coupling strength for a given synchronizability, reminiscent of the known linear relation between the noise intensity and the critical coupling strength $\epsilon_c = 2(D + \lambda)$ (as in our study we consider identical oscillators, we have $\lambda = 0$) for the mean-field Kuramoto model in the thermodynamic limit (Bag et al., 2007).

It is left to future work to assess whether this approximation can be derived analytically, at least for the mean-field Kuramoto model, and if it holds true in the case of non-identical oscillator models. Other issues to explore would be the influence of the oscillator model, the coupling method and the network topology on the parameters of the regression model.

Chapter 7. Conclusion

7.1. Contributions of this Thesis

This dissertation's primary focus is of a theoretical nature: a network-based framework capable of representing heterogeneous complex systems across scales. In combination with the accompanying software package, this framework, *deep graphs*, constitutes an important contribution to the field of complex systems, and possibly to data analytics in general. First, because of the explicit association of potentially unstructured, diverse information with the different (super)nodes and (super)edges, and second, because - in contrast to previous frameworks in the scientific literature - the properties of groups of objects (supernodes) and their respective interrelations and interactions (superedges) are incorporated into a self-contained network representation. For these reasons, our framework bridges the gap between (big) data and its modelling, and is capable of acting as a go-between, joining a unified and generalized network representation of systems with the tools and methods of fields such as multivariate statistics, probability theory and statistical physics, as well as the rising field of machine learning.

We presented a number of applications, demonstrating benefits of the deep graph framework: In an explorative analysis of global extreme rainfall measurements, we constructed a deep graph to track and categorize the formation of spatio-temporal rainfall clusters. We found propagation patterns over subtropical South America that were just recently discovered using rather complicated statistical methods, as well as extreme rainfall clusters over tropical South America that have not yet been identified and analyzed in the meteorological literature. Based on the constructed rainfall deep graph, we could also provide statistical evidence that the spatio-temporally integrated size distribution of extreme rainfall clusters does not - as previously suggested - follow a powerlaw. Instead, we found that the size distribution over the oceans is best approximated by an exponentially truncated powerlaw. Arguing with a generative storm-track model, we found that the exponential truncation of the observed distribution could be caused by the presence of land masses. In another application of the deep graph framework, we combined two high-resolution satellite products in order to identify spatio-temporal clusters of fire-affected areas in the Brazilian Amazon and characterize their land use specific burning conditions. By means of the revealed statistical characteristics, we took the first steps towards a probabilistic classifier of fire clusters into land use types with the ultimate goal of predicting whether a measured fire cluster was caused by anthropogenic activities or natural causes. Finally, we investigated the effects of white noise and global coupling strength on the maximum degree of synchronization for a variety of oscillator models in a broad spectrum of network topologies. We found a general sigmoidal scaling and validated it with a suitable regression model.

More detailed conclusions are drawn in the following.

Deep Graphs

We have introduced a collection of definitions resulting in *deep graphs*, a network-based framework enabling a mathematically accurate description of any given system in a self-contained manner. Our framework unifies existing network representations and generalizes them by fulfilling two essential objectives: a comprehensive treatment of groups of objects (supernodes) and their respective interrelations (superdeges), and an explicit association of information with the (super)nodes and (super)edges of a graph. The latter objective is implemented by specifying the (super)nodes and (super)edges of a (super)graph as sets of their respective properties. The former objective is implemented by transferring the mathematical concept of partition lattices to our graph representation. A deep graph, by our definition, is the set of all possible partitions of a graph, i.e. the partition lattice of a graph. The partition lattice of the edge set of a given graph is generally not covered entirely by the lattice of the graph, and can be queried on its own to gain insightful information (see Sec. 2.4.5). Together, the implemented objectives make it possible to aggregate, compute and allocate information on and between arbitrary groups of nodes. This information can then be stored on the lattices of a graph, allowing us to express and study properties, relations and interactions on all scales of the represented system(s).

In addition to its descriptive benefits, we were able to show how deep graphs establish an interface between graph theory, traditional data analysis and modelling tools, and machine learning methods. Furthermore, we introduced additional tools to support a comprehensive data analysis. The auxiliary *connector* and *selector* functions (see Sec. 2.5) make it easy to create and filter (super)edges, thereby allowing us to forge the topology of a deep graph. We have demonstrated how the properties of the nodes of a given graph induce a certain subset of the graph's partition lattice. When limiting ourselves to the representation of a multilayer network, this subset of the partition lattice corresponds to the different representations of the multilayer network, from its supra-graph representation to a tensor-like representation (see Appendix D). The concept of intersection partitions (see Sec. 2.4.5) allows us to calculate similarity measures between partitions of a graph (see Appendix A), and to express elaborate queries on the information contained in a deep graph.

To utilize the benefits of our framework, we provide a software implementation (the *DeepGraph Python Package*) that integrates seamlessly into the *PyData Ecosystem*, making it accessible to a large number of computational scientist.

Associated publication: Traxl et al., Chaos (2016, P2).

Tracking and Clustering of Extreme Rainfall

Employing the deep graphs framework, we conducted an explorative analysis of extreme rainfall events derived from a global, high-resolution satellite product. We first represented extreme rainfall events as nodes of a graph, whose features indicate their location, time and rainfall rate. We created edges between nodes reflecting the spatio-temporal proximity of the events, and then identified cohesive rainfall clusters as the connected components of the graph. This allowed us to track and visualize the clusters' temporal resolution, and calculate a number of characterizing features: their lifetime, their spatial coverage, and the total volume of water they precipitated. We further coarse-grained rainfall clusters into regional families, based on a measure of spatial overlap between them. We have discussed climatological characteristics of two of these families over the South American continent. The first family, concentrated over the subtropics, was just recently discovered using a rather complicated statistical methodology. The second, concentrated over tropical South America, has to our knowledge not yet been identified and analyzed in the meteorological literature.

Based on this explorative analysis, we could furthermore provide evidence that the spatio-temporally integrated size distribution of extreme rainfall clusters does not - as previously suggested - follow a powerlaw. Instead, we found that the distribution of rainfall clusters over the oceans is best approximated by an exponentially truncated powerlaw. Motivated by the fact that the conditions for strong cyclogenesis are typically not met over land, we hypothesized that the distribution could, in principle, follow a scale-free distribution on a planet without land masses, and that the exponential truncation is caused by the presence of land masses. To test this hypothesis, we proposed a generative model of synthetic storms with powerlaw-distributed lifetimes, evolving in a finite spatial area with absorbing boundaries. This simple model reproduces the exponentially truncated powerlaw observed for extreme rainfall clusters over the oceans, indicating that the proposed hypothesis suffices to explain the distributional characteristics discovered here.

Associated publications: Traxl et al., Chaos (2016, P2), Traxl et al., Geophysical Research Letters (accepted, P3).

Fire-Cluster Burning Conditions in the Amazonian Ecosystem

We advanced the understanding to what extent different land use types influence fire occurrence in the Amazonian ecosystem, which is particularly relevant for its conservation. We first combined two high-resolution satellite products - maps of fire-affected areas and land cover maps showing a detailed land use classification - into a graph representation, where nodes correspond to active fire measurements, and edges represent the spatio-temporal proximity between pairs of active fires. Similar to the analysis of rainfall measurements, we identified cohesive spatio-temporal fire clusters as the connected components of the graph, and calculated a set of features for each of them: their diameter; their lifetime; their "type", given by the predominant land use type a cluster occurred on; and their "first land use types", given by the set of land use types fires of a given cluster occurred on at the first satellite pass.

Investigating the distributions of diameters and lifetimes for the different cluster types, we found that savannah-type clusters dominate in terms of frequency, but also largest sizes (more than 30 km diameter) and longest lifetimes (more than 100 hours), followed by the land use types "Forest", "Pasture", "Agriculture", "Secondary Vegetation" and "Deforested", which show a consistent decrease in frequency and slopes. According to a maximum-likelihood evaluation, all distributions exhibit heavy tails, i.e. their tails are not exponentially bounded.

Regarding the originating land use types of "Mixed" clusters (i.e., clusters that propagated along various land use types and thus can not be assigned a dominant land use type), we found that 43% have "Pasture", 9% have "Agriculture", and 7% have "Deforested" in their first land use types. This indicates an overrepresentation of these land use types with respect to their overall proportions, which is particularly relevant in view of the fact that fires on these land use types are frequently caused by anthropogenic activities.

Finally, we derived a probabilistic classifier of fire clusters into dominant land use types, based on the clusters' diameters and lifetimes. We found that large clusters burning for short times are more likely to be savanna-type than "Mixed", which means their propagation velocity, on average, exceeds that of "Mixed" clusters. The other cluster types are predominantly small and short-lasting. Overall, the best guess for any given diameter/lifetime combination is either savannah-type or "Mixed". Discarding "Mixed" clusters, however, two small islands arise in the decision matrix (Fig. 5.7), showing that for some diameter/lifetime combinations, "Pasture" is more likely than savannah-type and for other combinations, "Forest" is more likely than savannah-type. The statistical significance of these islands, however, has yet to be determined.

Associated publication: Traxl et al., in preparation (P4)

Synchronizability in Noisy Complex Networks

We performed numerical simulations of a variety of oscillator models (Fitzhugh-Nagumo, Izhikevich and Kuramoto) coupled by both linear and non-linear coupling functions according to a broad spectrum of network topologies (scale-free, small-world, regular, random and highly modular). We were interested in the maximum degree of synchronization in dependence on the global coupling strength and the intensity of the white noise sources, and found common characteristics independent of the oscillator model, network and coupling type.

We have shown that the maximum degree of synchronization can be approximated quite adequately by a simple two-dimensional sigmoidal function with a linear relation between noise intensity and global coupling strength for a given synchronizability, reminiscent of the known linear relation between the noise intensity and the critical coupling strength $\epsilon_c = 2(D + \lambda)$ for the mean-field Kuramoto model in the thermodynamic limit.

Associated publication: Traxl et al., New Journal of Physics (2014, P1)

7.2. Outlook

Although the deep graph framework proposed in this dissertation will likely result in important applications in different areas of complex networks, we hope that it also initiates attempts to generalize existing network-specific measures and to develop new ones. Of particular interest would be measures related to the heterogeneity of a system's components and their interactions on different scales. In the context of multilayer networks, generalizations of network measures have already led to significant new insights, and we expect the same to become true for deep graphs.

Considering the explorative analysis of extreme rainfall events (see chapter 3), so far we have only investigated two regional families of rainfall clusters. It would not be surprising if a number of yet unknown propagation patterns are found in other families, given further inspection. An interesting extension of the study would be to measure the temporal similarities between families to investigate potential large-scale synchronizations between them.

Despite the fact that our approach to investigate fire cluster burning conditions (see chapter 5) is novel and preliminary results seem promising, some work is still left to be conducted and a few ideas are worth incorporating into the analysis in the future. Primarily, we need to fine-tune the spatio-temporal distance thresholds and validate them with the help of additional information. We could, for instance, compare the diameter and lifetime distributions with other datasets, such as the burned-area distributions analyzed by Cano-Crespo et al. (2015). Regarding the probabilistic classifier of fire clusters into land use types, a number of considerations to improve its predictive power come to mind. We could, for instance, look for differences in the burning profiles of the different cluster types in the dry and wet seasons. The recurrence rate of fires on a given spatial location could help us identify pasture fields, since farmers often burn their fields repeatedly to re-new the grass for their cattle. The study can also easily be extended to other regions on the planet.

Ultimately, a very promising project would be to combine different climatological data (such as the rainfall, fire and land use type data, together with lightning data) into one deep graph representation. This could allow us, for instance, to match fire clusters with rainfall and lightning data, and thereby improve our probabilistic classifier of fire clusters into land use types.

Appendix

Appendix A.

Measuring the Similarity of (Intersection) Partitions

We demonstrate how the construction of intersection partitions provides us with the elements of a so-called confusion matrix (or contingency table). These are necessary to compute similarity measures between partitions, such as, e.g.: the Jaccard index (Jaccard, 1912); the normalized mutual information (Strehl and Ghosh, 2003); or the normalized variation of information metric (Meilă, 2007). First, we show how to compute the similarity of two "normal" partitions, and then how to compute the similarity of two intersection partitions.

Assume we are given a graph $G = (V, E)$ comprised of n nodes, and two partitions of the node set, $V^p = \{V_i^p \,\big|\, i = 1, 2, ..., n^p\}$ and $V^{p'} = \{V_{i'}^{p'} \,\big|\, i' = 1, 2, ..., n^{p'}\}$. The number of nodes in supernode V_i^p ($V_{i'}^{p'}$) is then given by $n^{p,i}$ ($n^{p',i'}$), and the number of nodes in supernode $V_{i \cdot i'}^{p \cdot p'}$ of the intersection partition $V^{p \cdot p'}$ is given by $n^{p \cdot p', i \cdot i'}$ [see Eqs. (2.27)-(2.29)]. With these numbers, we can calculate the normalized variation of information metric by

$$\mathrm{NVI} = \frac{-1}{\log n} \sum_i \sum_{i'} \frac{n^{p \cdot p', i \cdot i'}}{n} \log \frac{(n^{p \cdot p', i \cdot i'})^2}{n^{p,i} n^{p',i'}}. \tag{A.1}$$

Analogously, we can compute other similarity measures, such as the Jaccard index or the normalized mutual information index (see Eqs. (6) and (7) in Granell et al. (2015)).

More generally, we can compute the similarity of two intersection partitions. Assume we are given a graph $G = (V, E)$ comprised of n nodes, and set of K partitions of V, induced by a set of functions ${}^v p = \{{}^v p^k \,|\, k \in I^K\}$, where $I^K = \{1, 2, ..., K\}$ is the partition index set. From this set of available partitions, we choose two collections, $g \subseteq I^K$ and $g' \subseteq I^K$, whose corresponding intersection partitions we want to compare. The number of nodes in supernode $V_{\underline{i}}^{\underline{p}}$ ($V_{\underline{i}'}^{\underline{p}'}$) is given by $n^{\underline{p},\underline{i}}$ ($n^{\underline{p}',\underline{i}'}$), and the number of nodes in supernode $V_{\underline{i} \cdot \underline{i}'}^{\underline{p} \cdot \underline{p}'}$ of the intersection partition $V^{\underline{p} \cdot \underline{p}'}$ is given by $n^{\underline{p} \cdot \underline{p}', \underline{i} \cdot \underline{i}'}$ (where $\underline{p} = (p^k)_{k \in g}$, $\underline{i} = (i^k)_{k \in g}$, $\underline{p}' = (p^k)_{k \in g'}$, $\underline{i}' = (i^k)_{k \in g'}$, and $i^k \in \{1, 2, ..., n^{p^k}\}$). Using these numbers in Eq. (A.1), we can compute the similarity of two different intersection partitions.

Equivalently, we can use the numbers $m, m^{\underline{p},\underline{ij},\underline{r}}, m^{\underline{p'},\underline{i'}\underline{j'},\underline{r'}}$ and $m^{\underline{pp'},\underline{ii'}\underline{jj'},\underline{rr'}}$ (see Tab. 2.1) to calculate similarity measures between (intersection) partitions of the edge set. Furthermore, we can use a pair of (intersection) partitions of the node set, in order to compute the similarity of their *corresponding* edge set partitions.

Appendix B.

Expressing Supernodes (Superedges) by Features (Relations)

We explicitly demonstrate how the information contained in a given graph $G = (V, E)$ is conserved when creating partitions, by expressing supernodes and superedges in terms of features and relations, respectively. Given a partition $V^{\underline{p}}$ of V induced by $\underline{p}$ [see Eqs. (2.27)-(2.29)], the set of features contained in supernode $V_{\underline{i}}^{\underline{p}}$ is given by

$$F_{\underline{i}}^{\underline{p}} = \{F_j^m \,\Big|\, j \in \{1, 2, ..., n\} \wedge m \in \{1, 2, ..., f_j\} \wedge \forall k \in g : {}^v p^k(V_j) = {}^v S_{i^k}^k\}. \tag{B.1}$$

To keep track of a features' original node index, and to guarantee uniqueness of every single feature, we technically would have to write (j, F_j^m) for every feature. Yet, for ease of notation, we refrain from doing so. Next, we map each feature F_j^m in $F_{\underline{i}}^{\underline{p}}$ onto its respective type,

$$t_{\underline{i}}^{\underline{p}} : F_{\underline{i}}^{\underline{p}} \to T_{\underline{i}}^{\underline{p}} = \{1, 2, ..., n_{\text{types}}^{\underline{p},\underline{i}}\}, F_j^m \mapsto t_{\underline{i}}^{\underline{p}}(F_j^m) := T_{\underline{i},t}^{\underline{p}} \in T_{\underline{i}}^{\underline{p}}, \tag{B.2}$$

such that $t_{\underline{i}}^{\underline{p}}(F_j^l) = t_{\underline{i}}^{\underline{p}}(F_k^m)$ for all pairs of features in $F_{\underline{i}}^{\underline{p}}$ that share the same type. We denote the number of distinct types of features in supernode $V_{\underline{i}}^{\underline{p}}$ by $n_{\text{types}}^{\underline{p},\underline{i}}$. Note that $0 \leq n_{\text{types}}^{\underline{p},\underline{i}} \leq |F_{\underline{i}}^{\underline{p}}|$, where $n_{\text{types}}^{\underline{p},\underline{i}} = 0$ either because the supernode $V_{\underline{i}}^{\underline{p}}$ does not exist, $n^{\underline{p},\underline{i}} = 0$, or because all the nodes it contains have no features, $n^{\underline{p},\underline{i}} \geq 1$ and $V_j = \{j\}$ for all $V_j \in V_{\underline{i}}^{\underline{p}}$. If no pair of nodes in $V_{\underline{i}}^{\underline{p}}$ shares any type of feature, then $n_{\text{types}}^{\underline{p},\underline{i}} = |F_{\underline{i}}^{\underline{p}}|$. The function $t_{\underline{i}}^{\underline{p}}$ induces a partition $F_{\underline{i}}^{\underline{p},T}$ of $F_{\underline{i}}^{\underline{p}}$ into features of common type $F_{\underline{i},t}^{\underline{p},T}$, given by

$$\begin{aligned} F_{\underline{i},t}^{\underline{p},T} = \{F_j^m \,\Big|\, & j \in \{1, 2, ..., n\} \wedge m \in \{1, 2, ..., f_j\} \wedge \\ & \forall k \in g : {}^v p^k(V_j) = {}^v S_{i^k}^k \wedge t_{\underline{i}}^{\underline{p}}(F_j^m) = T_{\underline{i},t}^{\underline{p}}\}, \end{aligned} \tag{B.3}$$

and $F_{\underline{i}}^{\underline{p},T} = \{F_{\underline{i},t}^{\underline{p},T} \,\Big|\, t \in \{1, 2, ..., n_{\text{types}}^{\underline{p},\underline{i}}\}\}$. We denote the number of features of type t in supernode $V_{\underline{i}}^{\underline{p}}$ by $n_{\text{t}}^{\underline{p},\underline{i}} := |F_{\underline{i},t}^{\underline{p},T}|$. Hence, we can express a supernode $V_{\underline{i}}^{\underline{p}}$ as a set of sets of features of common type (and its index, to guarantee uniqueness of the supernodes),

$$V^{\underline{p}}_{\underline{i}} = \{\underline{i}\} \cup \{F^{\underline{p},T}_{\underline{i},t}\}_{t\in\{1,2,\dots,n^{\underline{p},\underline{i}}_{\text{types}}\}}. \tag{B.4}$$

Analogously, we can express superedges in terms of their edges' constituent relations. Given a partition $E^{\underline{p}}$ of E induced by $\underline{p}$ [see Eqs. (2.33)-(2.40)], the set of relations contained in superedge $E^{\underline{p}}_{\underline{ij},\underline{r}}$ is given by

$$R^{\underline{p}}_{\underline{ij},\underline{r}} = \{R^m_{uv} \,\Big|\, \Phi^e(u,v) \wedge \Phi^v_{g^s}(u) \wedge \Phi^v_{g^t}(v) \wedge \Phi^e_{g^r}(u,v) \wedge m \in \{1,2,...,r_{uv}\}\}. \tag{B.5}$$

Again, to keep track of a relations' original indices and to guarantee uniqueness, we technically have to write $((u,v), R^m_{uv})$ for every relation, which we omit for notational clarity. Next, we map every relation R^m_{uv} in $R^{\underline{p}}_{\underline{ij},\underline{r}}$ onto its respective type,

$$t^{\underline{p}}_{\underline{ij},\underline{r}} : R^{\underline{p}}_{\underline{ij},\underline{r}} \to T^{\underline{p}}_{\underline{ij},\underline{r}} = \{1,2,...,m^{\underline{p},\underline{ij},\underline{r}}_{\text{types}}\}, R^m_{uv} \mapsto t^{\underline{p}}_{\underline{ij},\underline{r}}(R^m_{uv}) := T^{\underline{p}}_{\underline{ij},\underline{r},t} \in T^{\underline{p}}_{\underline{ij},\underline{r}}, \tag{B.6}$$

such that $t^{\underline{p}}_{\underline{ij},\underline{r}}(R^m_{ij}) = t^{\underline{p}}_{\underline{ij},\underline{r}}(R^n_{kl})$ for all pairs of relations in $R^{\underline{p}}_{\underline{ij},\underline{r}}$ that share the same type. We denote the number of distinct types of relations in superedge $E^{\underline{p}}_{\underline{ij},\underline{r}}$ by $m^{\underline{p},\underline{ij},\underline{r}}_{\text{types}}$. Again, $0 \leq m^{\underline{p},\underline{ij},\underline{r}}_{\text{types}} \leq |R^{\underline{p}}_{\underline{ij}}|$, where $m^{\underline{p},\underline{ij},\underline{r}}_{\text{types}} = |R^{\underline{p}}_{\underline{ij}}|$ only if no pair of edges in $E^{\underline{p}}_{\underline{ij},\underline{r}}$ shares any type of relation. The partition $R^{\underline{p},T}_{\underline{ij},\underline{r}}$ of $R^{\underline{p}}_{\underline{ij},\underline{r}}$ into relations of common type $R^{\underline{p},T}_{\underline{ij},\underline{r},t}$ is therefore induced by the function $t^{\underline{p}}_{\underline{ij},\underline{r}}$, where

$$\begin{aligned} R^{\underline{p},T}_{\underline{ij},\underline{r},t} = \{R^m_{uv} \,\Big|\, & \Phi^e(u,v) \wedge \Phi^v_{g^s}(u) \wedge \Phi^v_{g^t}(v) \wedge \Phi^e_{g^r}(u,v) \wedge \\ & m \in \{1,2,...,r_{uv}\} \wedge t^{\underline{p}}_{\underline{ij},\underline{r}}(R^m_{uv}) = T^{\underline{p}}_{\underline{ij},\underline{r},t}\}, \end{aligned} \tag{B.7}$$

and $R^{\underline{p},T}_{\underline{ij},\underline{r}} = \{R^{\underline{p},T}_{\underline{ij},\underline{r},t} \,\Big|\, t \in \{1,2,...,m^{\underline{p},\underline{ij},\underline{r}}_{\text{types}}\}\}$. We denote the number of relations of type t in superedge $E^{\underline{p}}_{\underline{ij},\underline{r}}$ by $m^{\underline{p},\underline{ij},\underline{r}}_{\mathrm{t}} := |R^{\underline{p},T}_{\underline{ij},\underline{r},t}|$. Therefore, a superedge $E^{\underline{p}}_{\underline{ij},\underline{r}}$ can be expressed as a set of sets of relations of common type (and its index, to guarantee uniqueness of the superedges),

$$E^{\underline{p}}_{\underline{ij},\underline{r}} = \{(\underline{i},\underline{j},\underline{r})\} \cup \{R^{\underline{p},T}_{\underline{ij},\underline{r},t}\}_{t\in\{1,2,...,m^{\underline{p},\underline{ij},\underline{r}}_{\text{types}}\}}. \tag{B.8}$$

Appendix C.

Summary of the Multilayer Network Representation

We summarize the representations of a multilayer network (MLN), as defined by Kivelä et al. (2014), and refer to the original paper for a more detailed description. A multilayer network (MLN) is defined by a quadruplet $M = (V_M, E_M, V^N, \boldsymbol{L})$, where the set of N nodes is given by $V^N = \{1, 2, ..., N\}$. The multidimensional layer structure is given by a sequence of sets of elementary layers, $\boldsymbol{L} = \{L_a\}_{a=1}^d$, where each of the d sets of elementary layers L_a corresponds to an 'aspect' a of the MLN (e.g., $L_1 = \{\text{facebook}, \text{twitter}, ...\}$ could be a set of categories of connections, and $L_2 = \{2010, 2011, ...\}$ could be a set of time stamps, at which edges are present). A layer in the structure given by $\boldsymbol{L}$ is then a combination of elementary layers from all aspects, or in other words: an element of the set of all layers given by the Cartesian product $L_1 \times \cdots \times L_d$. Each node can belong to any subset of the layers, and the set of all existing node-layer tuples (in short: node-layers) $(u, \boldsymbol{\alpha})$, where $u \in V^N$ and $\boldsymbol{\alpha} \in L_1 \times \cdots \times L_d$, is denoted $V_M \subseteq V^N \times L_1 \cdots \times L_d$. Edges are allowed between all such existing node-layers, hence the set of edges is given by $E_M \subseteq V_M \times V_M$.

The pair $G_M = (V_M, E_M)$, referred to as the 'supra-graph' of M, is a graph on its own, where nodes are, as the authors say, "labelled in a certain way". The adjacency matrix of G_M is referred to as the 'supra-adjacency matrix' representation of M, and constitutes one possible representation of a MLN. Defining weights for edges of M on the underlying graph G_M (by some function $w : E_M \to \mathbb{R}$) yields a weighted MLN.

Another representation of a MLN can be achieved by adjacency tensors (De Domenico et al., 2013). Given a MLN $M = (V_M, E_M, V^N, \boldsymbol{L})$ with d aspects, one can represent it by an order-$2(d+1)$ adjacency tensor $A_{uv\boldsymbol{\alpha\beta}} = A_{uv\alpha_1\beta_1...\alpha_d\beta_d}$, where an element $A_{uv\boldsymbol{\alpha\beta}}$ has a value of 1, if and only if $((u, \boldsymbol{\alpha}), (v, \boldsymbol{\beta})) \in E_M$, and a value of 0 otherwise. As Kivelä et al. (2014) explain, the representation of a MLN by an adjacency tensor is technically only valid for node-aligned MLNs, where all layers contain all nodes, $V_M = V^N \times L_1 \times \cdots \times L_d$. Yet, many tensor-based methods on MLNs have been successfully applied by filling layers with 'empty' node-layers (node-layers that are not adjacent to any other node-layer), yielding an artificial node-aligned structure of the MLN. However, one has to be very cautious in the calculation and interpretation of tensor-based measures, and account for the presence of empty node-layers in an appropriate way (Kivelä et al., 2014). In the tensor-representation of MLNs, weights can be introduced by defining a weighted adjacency tensor $W_{uv\boldsymbol{\alpha\beta}}$, where the value of each element determines the weight of an edge (for non-existing edges, the value is 0 by convention).

Appendix D. Discussion of Multilayer Networks

We first demonstrate the alternative representation of a multilayer network (MLN) by our framework, which is given by placing the additional information attributed to the layered structure of a MLN M in the edges of $G = (V, E)$. Then, we show the advantages of the representation stated in the main text. For that matter, we create the subset of the partition lattice ${}^{G}L$ of $G \mathrel{\widehat{=}} M$ that is induced by the types of features of its constituent nodes, and show that it incorporates not only the alternative representation shown here, but several others, including a tensor-like representation. Lastly, we discuss the constraints imposed on our framework in order to represent a MLN, and explain how our framework solves the issues encountered in the representation of MLNs.

The alternative representation of $M = (V_M, E_M, V^N, \boldsymbol{L})$ by $G = (V, E)$ is given by identifying each node $V_i = \{i\} \in V = \{V_1, V_2, ..., V_N\}$ with a node $V_i^N \in V^N$, $V_i \mathrel{\widehat{=}} V_i^N$. Denoting the weight of an edge of a MLN by $w\left(((V_i^N, \boldsymbol{\alpha}), (V_j^N, \boldsymbol{\beta}))\right) \in \mathbb{R}$, an edge $E_{ij} \in E' = \{E_{11}, E_{12}, ..., E_{NN}\}$ is given by

$$\begin{aligned} E_{ij} = \{w\left(((V_i^N, \boldsymbol{\alpha}), (V_j^N, \boldsymbol{\beta}))\right) \,\big|\, ((V_i^N, \boldsymbol{\alpha}), (V_j^N, \boldsymbol{\beta})) \in E_M\} \\ =: \{R_{ij}^k \,\big|\, k \in \{1, 2, ..., r_{ij}\}\}, \end{aligned} \tag{D.1}$$

where $|E_{ij}| = r_{ij}$ is the number of types of relations from node V_i to node V_j. Hence, the edge set E corresponding to E_M is given by $E = \{E_{ij} \,|\, i, j \in \{1, 2, ..., N\} \wedge E_{ij} \neq \varnothing\}$. By this representation, we can clearly see that a tuple $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ defines the type of relation of an edge in E_M,

$$t\left(((V_i^N, \boldsymbol{\alpha}), (V_j^N, \boldsymbol{\beta}))\right) = t\left(((V_k^N, \boldsymbol{\gamma}), (V_l^N, \boldsymbol{\delta}))\right) \longleftrightarrow (\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\boldsymbol{\gamma}, \boldsymbol{\delta}), \tag{D.2}$$

for all $V_i^N, V_j^N, V_k^N, V_l^N \in V^N$, where t is a function mapping an edge to its corresponding type, $t : E_M \to T = \{1, 2, ..., m_{\text{types}}\}$, with $m_{\text{types}} = (\prod_{a=1}^{d} |L_a|)^2$. Therefore, the number of types of relations between any pair of nodes in a MLN is bounded by $r_{ij} \leq m_{\text{types}}$.

Next, we partition the graph $G = (V, E) \mathrel{\widehat{=}} (V_M, E_M)$ described by Eqs. (2.49) and (2.50). For notational uniformity, we rewrite the features of the nodes in V as outputs of partition functions $p = \{p^N, p^1, p^2, ..., p^d\}$, where

$$p^N : V \to V^N, V_i \mapsto p^N(V_i) = V_i^N, \tag{D.3}$$

$$p^a : V \to L_a, V_i \mapsto p^a(V_i) = L_{a,i}, a = 1, 2, ..., d. \tag{D.4}$$

Based on the $(1 + d)$ partitions induced by p, we can redistribute the information contained in the graph G on a subset of the lattice ${}^{G}L^{f} \subseteq {}^{G}L$. This redistribution allows for several representations of the graph G, some of which we will demonstrate in the following. Let us denote the partition index set of p by $I^K = \{N, 1, 2, ..., d\}$. Then we can select a total of $I(K) = 2^{(1+d)}$ distinct collections $g \subseteq I^K$, resulting in $|{}^{G}L^{f}| \leq I(K)$ supergraphs $G^{\underline{p}} = (V^{\underline{p}}, E^{\underline{p}}) \in {}^{G}L^{f}$, where ${}^{G}L^{f} = \{G^{\underline{p}} \,|\, g \in \mathcal{P}(I^K)\}$ and $\underline{p} = (p^k)_{k \in g}$.

Choosing $g = \{N\}$ leads to the supergraph $G^{p^N} = (V^{p^N}, E^{p^N})$, where each supernode $V_i^{p^N} \in V^{p^N}$ corresponds to a node of the MLN, $V_i^{p^N} \mathrel{\widehat{=}} V_i^N$. Superedges $E_{ij}^{p^N} \in E^{p^N}$ with $i = j$ correspond to the coupling edges of a MLN. The one to one correspondence of the supergraph G^{p^N} to the above, edge-based choice of G justifies the statement that the representation G of M given in the main text is the preferred one, since it fully entails the above choice.

Choosing the group $g = \{1, 2, ..., d\}$ leads to the supergraph $G^{p^1 \cdots p^d} = (V^{p^1 \cdots p^d}, E^{p^1 \cdots p^d})$, where every supernode $V_{i^1 \cdots i^d}^{p^1 \cdots p^d} \in V^{p^1 \cdots p^d}$ corresponds to a distinct layer of M, encompassing all its respective nodes. Superedges $E_{i^1 \cdots i^d, j^1 \cdots j^d}^{p^1 \cdots p^d} \in E^{p^1 \cdots p^d}$ with either $(i^a)_{a=1}^d = (j^a)_{a=1}^d$ or $(i^a)_{a=1}^d \neq (j^a)_{a=1}^d$ correspond to collections of intra- and interlayer edges of a MLN, respectively.

The last supergraph we want to exemplify is given by choosing $g = \{N, 1, 2, ..., d\} = I^K$, resulting in the supergraph $G^{p^N \cdot p^1 \cdots p^d} = (V^{p^N \cdot p^1 \cdots p^d}, E^{p^N \cdot p^1 \cdots p^d})$. This supergraph corresponds one to one to the graph $G = (V, E)$, and therefore to the 'supra-graph' representation of M, given by the tuple (V_M, E_M). The only difference is the indexing. The graph G has an adjacency matrix-like representation, given by $E_{ij} \in E'$. We say 'like', since E' is not a matrix, formally. An element of E' is either a real number, corresponding to the weight of the corresponding edge in E_M, or an empty set, meaning the edge does not exist. $G^{p^N \cdot p^1 \cdots p^d}$, on the other hand, has a tensor-like representation, given by $E_{i^N \cdot i^1 \cdots i^d, j^N \cdot j^1 \cdots j^d}^{p^N \cdot p^1 \cdots p^d} \in E^{p^N \cdot p^1 \cdots p^d}$. Again, formally, $E^{p^N \cdot p^1 \cdots p^d}$ is not a tensor. An element of $E^{p^N \cdot p^1 \cdots p^d}$ is either a real number, corresponding to the weight of the corresponding edge in E_M, or an empty set, if the edge does not exist. As mentioned in Sec. 2.4.5, we can distinguish between a superedge that does not exist because at least one of the supernodes does not exist, $n^{p^N \cdot p^1 \cdots p^d, i^N \cdot i^1 \cdots i^d}$ or $n^{p^N \cdot p^1 \cdots p^d, j^N \cdot j^1 \cdots j^d} = 0$, or because there is no superedge between existing supernodes, $n^{p^N \cdot p^1 \cdots p^d, i^N \cdot i^1 \cdots i^d}$ and $n^{p^N \cdot p^1 \cdots p^d, j^N \cdot j^1 \cdots j^d} = 1$.

From the perspective of our framework, all representations $G^{\underline{p}} \in {}^{G}L$ are equivalent, in the sense that the information contained in G is conserved under partitioning. There is no need to "flatten" the MLN represented by $G^{p^N \cdot p^1 \cdots p^d}$ to obtain its supra-adjacency matrix representation G, and there is no loss of information about the aspects, as – according to Kivelä et al. (2014) – it is the case for MLNs represented by $M = (V_M, E_M, V^N, \boldsymbol{L})$.

Let us now summarize the constraints we imposed on our framework, in order to represent a MLN. First, we had to restrict ourselves to the representation of one element of a deep graph. Allocating information on and between groups of nodes, as described in Sec. 2.4.6, is not intended within the framework of MLNs. Then, we have to decide whether to put to information attributed to the layered structure of M into the nodes of G, or the edges of G. There is no genuine separation of features and relations in a MLN. Furthermore, the weights of the edges of a MLN need to be restricted to real numbers (or possibly complex numbers). This poses several limitations. First, it is problematic to distinguish between edges with a weight of 0 (e.g. an edge representing a time difference of 0) and non-existing edges, since edges with weight 0 do not exist by convention in MLNs. Yet, more importantly, we can not assign distributions of values to nodes or edges, let alone more complex mathematical objects. Another complication arises, when dealing with nodes that have more or less than d aspects, or more generally speaking: when dealing with heterogeneous kinds of nodes. Although it is possible to represent nodes with different types of features by filling layers with 'empty' node-layers, the procedure is rather counter-intuitive and leads to a cluttered representation. In contrast, our framework provides the means to represent heterogeneous objects and their relations in a sparse and intuitive manner.

Bibliography

Acebrón, J. A., L. L. Bonilla, C. J. Pérez-Vicente, F. Ritort, and R. Spigler (2005). "The Kuramoto model: A simple paradigm for synchronization phenomena". In: *Rev. Mod. Phys.* 77, pp. 137–185.

Aicher, Christopher, Abigail Z Jacobs, and Aaron Clauset (2013). "Adapting the stochastic block model to edge-weighted networks". In: *arXiv preprint arXiv:1305.5782*.

Almeida, Cláudio Aparecido de, Taıse Farias Pinheiro, Alda Monteiro Barbosa, Maria Rafaela Braga Salum de Abreu, Felipe de Lucia Lobo, Maurıcio Silva, Alessandra Rodrigues Gomes, Luis Waldir Rodrigues Sadeck, Lariana Têka Barra de Medeiros, Murilo Figueira Neves, et al. (2009). "Metodologia para mapeamento de vegetação secundária na Amazônia Legal". In:

Anabor, Vagner, David J. Stensrud, and Osvaldo L. L. de Moraes (2008). "Serial Upstream-Propagating Mesoscale Convective System Events over Southeastern South America". In: *Monthly Weather Review* 136.8, pp. 3087–3105. ISSN: 0027-0644. DOI: `10.1175/2007MWR2334.1`.

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471360919.

Andrade, R.F.S., H.J. Schellnhuber, and M. Claussen (1998). "Analysis of rainfall records: possible relation to self-organized criticality". In: *Physica A: Statistical Mechanics and its Applications* 254.3-4, pp. 557–568. ISSN: 03784371. DOI: `10.1016/S0378-4371(98)00057-0`.

Arenas, Alex, Albert Díaz-Guilera, and Conrad J. Pérez-Vicente (2006). "Synchronization Reveals Topological Scales in Complex Networks". In: *Phys. Rev. Lett.* 96 (11), p. 114102. DOI: `10.1103/PhysRevLett.96.114102`.

Armenteras, Dolors and Javier Retana (2012). "Dynamics, Patterns and Causes of Fires in Northwestern Amazonia". In: *PLoS ONE* 7.4, pp. 1–7. DOI: `10.1371/journal.pone.0035288`.

Arthur, David and Sergei Vassilvitskii (2007). "k-means++: The advantages of careful seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, pp. 1027–1035.

Asner, Gregory P, David E Knapp, Eben N Broadbent, Paulo JC Oliveira, Michael Keller, and Jose N Silva (2005). "Selective logging in the Brazilian Amazon". In: *Science* 310.5747, pp. 480–482.

Asratian, A.S., T.M.J. Denley, and R. Häggkvist (1998). *Bipartite Graphs and Their Applications*. Cambridge Tracts in Mathematics. Cambridge University Press. ISBN: 9780521593458.

Bag, Bidhan Chandra, K. G. Petrosyan, and Chin-Kun Hu (2007). "Influence of noise on the synchronization of the stochastic Kuramoto model". In: *Phys. Rev. E* 76 (5), p. 056210. DOI: 10.1103/PhysRevE.76.056210.

Barahona, Mauricio and L. M. Pecora (2002). "Synchronization in Small-World Systems". In: *Phys. Rev. Lett.* 89.5, p. 054101. DOI: 10.1103/physrevlett.89.054101.

Barona, Elizabeth, Navin Ramankutty, Glenn Hyman, and Oliver T Coomes (2010). "The role of pasture and soybean in deforestation of the Brazilian Amazon". In: *Environmental Research Letters* 5.2, p. 024002.

Bartsch, Ronny P. and Plamen Ch. Ivanov (2014). "Coexisting Forms of Coupling and Phase-Transitions in Physiological Networks". English. In: *Nonlinear Dynamics of Electronic Systems.* Ed. by Valeri M. Mladenov and Plamen Ch. Ivanov. Vol. 438. Communications in Computer and Information Science. Springer International Publishing, pp. 270–287. ISBN: 978-3-319-08671-2. DOI: 10.1007/978-3-319-08672-9_33.

Bashan, Amir, Ronny P. Bartsch, Jan W. Kantelhardt, Shlomo Havlin, and Plamen Ch Ivanov (2012). "Network physiology reveals relations between network topology and physiological function". In: *Nat Commun* 3, p. 702. DOI: 10.1038/ncomms1705.

Becker, H. W. and John Riordan (1948). "The Arithmetic of Bell and Stirling Numbers". English. In: *American Journal of Mathematics* 70.2, pp. 385–394. ISSN: 00029327.

Bell, E. T. (1938). "The Iterated Exponential Integers". English. In: *Annals of Mathematics.* Second Series 39.3, pp. 539–557. ISSN: 0003486X.

Belykh, Igor, Enno de Lange, and Martin Hasler (2005). "Synchronization of Bursting Neurons: What Matters in the Network Topology". In: *Physical Review Letters* 94.18, pp. 188101+. DOI: 10.1103/PhysRevLett.94.188101.

Berge, C. (1976). *Graphs and Hypergraphs.* Graphs and Hypergraphs. North-Holland Publishing Company. ISBN: 9780720424539.

Berlingerio, Michele, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino Pedreschi (2011). "Foundations of multidimensional network analysis". In: *Proceedings - 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011* June 2015, pp. 485–489. DOI: 10.1109/ASONAM.2011.103.

Birkhoff, Garrett (1940). *Lattice Theory, Volume 25, Part 2.* American Mathematical Soc., p. 418. ISBN: 0821810251.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Information Science and Statistics. Springer. ISBN: 9780387310732.

Boccaletti, Stefano, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesus Gómez-Gardeñes, Miguel Romance, Irene Sendiña-Nadal, Zhen Wang, and Massimiliano Zanin (2014). "The structure and dynamics of multilayer networks". In: *Physics Reports* 544.1, pp. 1–122. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2014.07.001.

Boers, Niklas, Bodo Bookhagen, Norbert Marwan, Jürgen Kurths, and José Marengo (2013). "Complex networks identify spatial patterns of extreme rainfall events of the South American Monsoon System". In: *Geophysical Research Letters* 40.16, pp. 4386–4392. ISSN: 00948276. DOI: 10.1002/grl.50681.

Boers, Niklas, Bodo Bookhagen, Henrique M. J. Barbosa, Norbert Marwan, Jürgen Kurths, and José Marengo (2014). "Prediction of Extreme Floods in the Eastern Central Andes based on a Complex Network Approach". In: *Nature Communications* 5:5199. DOI: `doi:10.1038/ncomms6199`.

Boers, Niklas, Henrique M. J. Barbosa, Bodo Bookhagen, José a. Marengo, Norbert Marwan, and Jürgen Kurths (2015). "Propagation of Strong Rainfall Events from Southeastern South America to the Central Andes". In: *Journal of Climate* 28.19, pp. 7641–7658. ISSN: 0894-8755. DOI: `10.1175/JCLI-D-15-0137.1`.

Bollobas, Bela (1998). *Modern Graph Theory*. Springer Science & Business Media, p. 394. ISBN: 1461206197.

Briegel, Lisa M and William M Frank (1997). "Large-Scale Influences on Tropical Cyclogenesis in the Western North Pacific". In: *Monthly Weather Review* 125, pp. 1397–1413. ISSN: 0027-0644. DOI: `10.1175/1520-0493(1997)125<1397:LSIOTC>2.0.CO;2`.

Buluç, Aydin, Henning Meyerhenke, Ilya Safro, Peter Sanders, and Christian Schulz (2013). "Recent Advances in Graph Partitioning". In: *CoRR* abs/1311.3144.

Cahalan, Robert F. and Joachim H. Joseph (1989). *Fractal Statistics of Cloud Fields*. DOI: `10.1175/1520-0493(1989)117<0261:FSOCF>2.0.CO;2`.

Campanharo, Andriana SLO, M Irmak Sirer, R Dean Malmgren, Fernando M Ramos, and Luís A Nunes Amaral (2011). "Duality between time series and networks". In: *PloS one* 6.8, e23378. DOI: `10.1371/journal.pone.0023378`.

Cano-Crespo, Ana, Paulo JC Oliveira, Alice Boit, Manoel Cardoso, and Kirsten Thonicke (2015). "Forest edge burning in the Brazilian Amazon promoted by escaping fires from managed pastures". In: *Journal of Geophysical Research: Biogeosciences* 120.10, pp. 2095–2107.

Chen, Sheng, Yang Hong, Jonathan J Gourley, George J Huffman, Yudong Tian, Qing Cao, Bin Yong, Pierre-Emmanuel Kirstetter, Junjun Hu, Jill Hardy, Zhe Li, Sadiq I Khan, and Xianwu Xue (2013). "Evaluation of the successive V6 and V7 TRMM multisatellite precipitation analysis over the Continental United States". In: *Water Resources Research* 49.12, pp. 8174–8186. ISSN: 1944-7973. DOI: `10.1002/2012WR012795`.

Cheng, Chee-Pong and Robert A. Houze Jr. (1979). "The Distribution of Convective and Mesoscale Precipitation in GATE Radar Echo Patterns". In: *Monthly Weather Review* 107.10, pp. 1370–1381. ISSN: 0027-0644. DOI: `10.1175/1520-0493(1979)107<1370:TDOCAM>2.0.CO;2`.

Clauset, A, CR Shalizi, and MEJ Newman (2009). "Power-law distributions in empirical data". In: *SIAM review* 51.4, pp. 661–703. arXiv: `arXiv:0706.1062v2`.

Cochrane, Mark A (2003). "Fire science for rainforests". In: *Nature* 421.6926, pp. 913–919.

Cochrane, Mark A and William F Laurance (2008). "Synergisms among fire, land use, and climate change in the Amazon". In: *AMBIO: A Journal of the Human Environment* 37.7, pp. 522–527.

Cochrane, Mark A, Ane Alencar, Mark D Schulze, Carlos M Souza, Daniel C Nepstad, Paul Lefebvre, and Eric A Davidson (1999). “Positive feedbacks in the fire dynamic of closed canopy tropical forests”. In: *Science* 284.5421, pp. 1832–1835.

Cohen, Júlia C. P., Maria A. F. Silva Silva Dias, and Carlos A. Nobre (1995). “Environmental Conditions Associated with Amazonian Squall Lines: A Case Study”. In: *Monthly Weather Review* 123.11, pp. 3163–3174.

Comaniciu, Dorin and Peter Meer (2002). “Mean shift: A robust approach toward feature space analysis”. In: *IEEE Transactions on pattern analysis and machine intelligence* 24.5, pp. 603–619.

Commons, Wikimedia (2015). *The 15 partitions of a 4-element set ordered in a Hasse diagram.* `https://commons.wikimedia.org/wiki/File:Set_partitions_4;_Hasse;_circles.svg`. Accessed: 2015-08-29.

Davidson, Eric A, Alessandro C de Araújo, Paulo Artaxo, Jennifer K Balch, I Foster Brown, Mercedes MC Bustamante, Michael T Coe, Ruth S DeFries, Michael Keller, Marcos Longo, et al. (2012). “The Amazon basin in transition”. In: *Nature* 481.7381, pp. 321–328.

De Domenico, Manlio, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A. Porter, Sergio Gómez, and Alex Arenas (2013). “Mathematical Formulation of Multilayer Networks”. In: *Physical Review X* 3.4, p. 041022. ISSN: 2160-3308. DOI: `10.1103/PhysRevX.3.041022`.

DeepGraph Python Package. `https://github.com/deepgraph/deepgraph/`. Accessed: 2016-03-30.

Deng, Li and Dong Yu (2014). *DEEP LEARNING: Methods and Applications.* Tech. rep. MSR-TR-2014-21.

Dickman, Ronald (2003). “Rain, power laws, and advection”. In: *Physical Review Letters* 90.10, p. 108701. ISSN: 0031-9007. DOI: `10.1103/PhysRevLett.90.108701`. arXiv: `0210327 [cond-mat]`.

Durkee, Joshua D. and Thomas L. Mote (2009). “A climatology of warm-season mesoscale convective complexes in subtropical South America”. In: *International Journal of Climatology* 30.3, pp. 418–431. DOI: `10.1002/joc.1893`.

Durkee, Joshua D., Thomas L. Mote, and J. Marshall Shepherd (2009). “The contribution of mesoscale convective complexes to rainfall across subtropical South America”. In: *Journal of Climate* 22.17, pp. 4590–4605. ISSN: 0894-8755. DOI: `10.1175/2009JCLI2858.1`.

Erdős, Paul and Alfréd Rényi (1959). “On random graphs”. In: *Publicationes Mathematicae Debrecen* 6, pp. 290–297.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. (1996). “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: *Kdd.* Vol. 96. 34, pp. 226–231.

Field, C B, V Barros, T F Stocker, D Qin, D J Dokken, K L Ebi, M D Mastrandrea, K J Mach, S K Allen, and M Tignor (2012a). “IPCC, 2012: Glossary of terms”. In: *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation - A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC)*, pp. 555–564. DOI: `10.14573/altex.140331`.

– (2012b). "IPCC, 2012: Glossary of terms". In: *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation - A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC)*, pp. 555–564. DOI: 10.1177/1403494813515131.

Fitzhugh, R. (1961). "Impulses and Physiological States in Theoretical Models of Nerve Membrane". In: *Biophysical Journal* 1.6, pp. 445–466. ISSN: 00063495. DOI: 10.1016/s0006-3495(61)86902-6.

Gao, Jianxi, Sergey V. Buldyrev, H. Eugene Stanley, and Shlomo Havlin (2011a). "Networks formed from interdependent networks". In: *Nature Physics* 8.1, pp. 40–48. ISSN: 1745-2473. DOI: 10.1038/nphys2180.

Gao, Jianxi, Sergey V. Buldyrev, Shlomo Havlin, and H. Eugene Stanley (2011b). "Robustness of a network of networks". In: *Physical Review Letters* 107.19, pp. 1–5. ISSN: 00319007. DOI: 10.1103/PhysRevLett.107.195701. arXiv: 1010.5829.

Giglio, Louis, Jacques Descloitres, Christopher O Justice, and Yoram J Kaufman (2003). "An enhanced contextual fire detection algorithm for MODIS". In: *Remote sensing of environment* 87.2, pp. 273–282.

Glass, Leon, Peter Hunter, Andrew McCulloch, and Institute for Nonlinear Science (1991). *Theory Of Heart: Biomechanics, Biophysics, And Nonlinear Dynamics Of Cardiac Function.* New York: Springer-Verlag. ISBN: 9780387974835.

Goldenberg, S. B. (2001). "The Recent Increase in Atlantic Hurricane Activity: Causes and Implications". In: *Science* 293.5529, pp. 474–479. ISSN: 00368075. DOI: 10.1126/science.1060040.

Gómez-Gardeñes, J., Y. Moreno, and A. Arenas (2007). "Synchronizability determined by coupling strengths and topology on Complex Networks". In: *Phys. Rev. E* 75, p. 066106.

Granell, Clara, Richard K. Darst, Alex Arenas, Santo Fortunato, and Sergio Gómez (2015). "Benchmark model to assess community structure in evolving networks". In: *Physical Review E* 92.1, p. 012805. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.92.012805.

Han, Jaiwei (2012). "Mining Heterogeneous Information Networks: The Next Frontier". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '12. Beijing, China: ACM, pp. 2–3. ISBN: 978-1-4503-1462-6. DOI: 10.1145/2339530.2339533.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer Series in Statistics. Springer. ISBN: 9780387848587.

Haykin, S.S. (2009). *Neural Networks and Learning Machines.* Neural networks and learning machines Bd. 10. Prentice Hall. ISBN: 9780131471399.

Horvath, S. (2014). *Weighted Network Analysis: Applications in Genomics and Systems Biology.* Springer New York. ISBN: 9781493900220.

Houze Jr., Robert A. and Chee-Pong Cheng (1977). *Radar Characteristics of Tropical Convection Observed During GATE: Mean Properties and Trends Over the Summer Season.* DOI: 10.1175/1520-0493(1977)105<0964:RCOTCO>2.0.CO;2.

Huffman, George J, David T Bolvin, Eric J Nelkin, David B Wolff, Robert F Adler, Guojun Gu, Yang Hong, Kenneth P Bowman, and Erich F Stocker (2007). “The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales”. In: *Journal of Hydrometeorology* 8.1, pp. 38–55.

Ivanov, Plamen Ch. and Ronny P. Bartsch (2014). “Network Physiology: Mapping Interactions Between Networks of Physiologic Networks”. English. In: *Networks of Networks: The Last Frontier of Complexity.* Ed. by Gregorio D’Agostino and Antonio Scala. Understanding Complex Systems. Springer International Publishing, pp. 203–222. ISBN: 978-3-319-03517-8. DOI: 10.1007/978-3-319-03518-5_10.

Izhikevich, E. M. (2003). “Simple model of spiking neurons”. In: *IEEE Transactions on Neural Networks* 14.6, pp. 1569–1572. ISSN: 1045-9227. DOI: 10.1109/TNN.2003.820440.

– (2004). “Which model to use for cortical spiking neurons?” In: *Neural Networks, IEEE Transactions on* 15.5, pp. 1063–1070. ISSN: 1045-9227. DOI: 10.1109/TNN.2004.832719.

Jaccard, Paul (1912). “The distribution of the flora in the alpine zone”. In: *New phytologist* 11.2, pp. 37–50. DOI: 10.1111/j.1469-8137.1912.tb05611.x.

Jaynes, E. T. (2005). “Probability theory: the logic of science”. In: *The Mathematical Intelligencer* 27.2, pp. 83–83. ISSN: 0343-6993. DOI: 10.1007/BF02985800.

Kivelä, Mikko, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter (2014). “Multilayer networks”. In: *Journal of Complex Networks* 2.3, pp. 203–271. DOI: 10.1093/comnet/cnu016.

Koch, C. (1999). *Biophysics of Computation: Information Processing in Single Neurons.* Computational neuroscience. Oxford University Press. ISBN: 9780195104912.

Kuramoto, Yoshiki (1984). *Chemical Oscillations, Waves, and Turbulence.* Vol. 19. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-69691-6. DOI: 10.1007/978-3-642-69689-3.

Laurance, William F (1999). “Reflections on the tropical deforestation crisis”. In: *Biological Conservation* 91.2, pp. 109–117.

Laurance, William F, Mark A Cochrane, Scott Bergen, Philip M Fearnside, Patricia Delamônica, Christopher Barber, Sammya D’Angelo, and Tito Fernandes (2001). “The future of the Brazilian Amazon”. In: *Science* 291.5503, pp. 438–439.

Lay, Erin H, Abram R Jacobson, Robert H Holzworth, Craig J Rodger, and Richard L Dowden (2007). “Local time variation in land/ocean lightning flash density as measured by the World Wide Lightning Location Network”. In: *Journal of Geophysical Research: Atmospheres* 112.D13.

Lewis, Simon L, Paulo M Brando, Oliver L Phillips, Geertje MF van der Heijden, and Daniel Nepstad (2011). “The 2010 amazon drought”. In: *Science* 331.6017, pp. 554–554.

Liu, Yongqiang, Scott Goodrick, and Warren Heilman (2014). “Wildland fire emissions, carbon, and climate: Wildfire–climate interactions”. In: *Forest Ecology and Management* 317, pp. 80–96.

López, R E (1977). *The Lognormal Distribution and Cumulus Cloud Populations.* DOI: 10.1175/1520-0493(1977)105<0865:TLDACC>2.0.CO;2.

Lucas, John F. (1990). *Introduction to Abstract Mathematics.* Rowman & Littlefield, p. 382. ISBN: 091267573X.

Maddox, Robert A. (1980). "Mesoscale convective complexes". In: *Bulletin of the American Meteorological Society* 61.11, pp. 1374–1387. DOI: 10.1175/1520-0477(1980)061<1374:MCC>2.0.CO;2.

Mapes, Brain E. and Robert A. Houze Jr. (1993). *Cloud Clusters and Superclusters over the Oceanic Warm Pool.* DOI: 10.1175/1520-0493(1993)121<1398:CCASOT>2.0.CO;2.

Marengo, J A, B Liebmann, A M Grimm, V Misra, P L Silva Dias, I F A Cavalcanti, L M V Carvalho, E H Berbery, T Ambrizzi, C S Vera, Others, P L Silva Dias, A C Saulo, J Nogues-paegle, E Zipser, A Seth, and L M Alves (2012). "Recent developments on the South American monsoon system". In: *International Journal of Climatology* 32.1, pp. 1–21. DOI: 10.1002/joc.2254.

Marwan, N. (2008). "A historical review of recurrence plots". English. In: *The European Physical Journal Special Topics* 164.1, pp. 3–12. ISSN: 1951-6355. DOI: 10.1140/epjst/e2008-00829-1.

Marwan, Norbert and Jürgen Kurths (2002). "Nonlinear analysis of bivariate data with cross recurrence plots". In: *Physics Letters, Section A: General, Atomic and Solid State Physics* 302.5-6, pp. 299–307. ISSN: 03759601. DOI: 10.1016/S0375-9601(02)01170-2.

Meilă, Marina (2007). "Comparing clusterings — an information based distance". In: *Journal of multivariate analysis* 98.5, pp. 873–895. DOI: 10.1016/j.jmva.2006.11.013.

Morton, Douglas C, Ruth S DeFries, Yosio E Shimabukuro, Liana O Anderson, Egidio Arai, Fernando del Bon Espirito-Santo, Ramon Freitas, and Jeff Morisette (2006). "Cropland expansion changes deforestation dynamics in the southern Brazilian Amazon". In: *Proceedings of the National Academy of Sciences* 103.39, pp. 14637–14641.

Motter, Adilson E., Changsong Zhou, and Jürgen Kurths (2005). "Network synchronization, diffusion, and the paradox of heterogeneity". In: *Phys. Rev. E* 71 (1), p. 016116. DOI: 10.1103/PhysRevE.71.016116.

Neggers, R. A. J., H. J. J. Jonker, and A. P. Siebesma (2003). "Size Statistics of Cumulus Cloud Populations in Large-Eddy Simulations". In: *Journal of the Atmospheric Sciences* 60.8, pp. 1060–1074. ISSN: 0022-4928. DOI: 10.1175/1520-0469(2003)60<1060:SSOCCP>2.0.CO;2.

Nesbitt, Stephen W., Robert Cifelli, and Steven a. Rutledge (2006). "Storm Morphology and Rainfall Characteristics of TRMM Precipitation Features". In: *Monthly Weather Review* 134.10, pp. 2702–2721. ISSN: 0027-0644. DOI: 10.1175/MWR3200.1.

Newman, M. and M. Girvan (2004). "Finding and evaluating community structure in networks". In: *Physical Review E* 69.2, p. 026113. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.69.026113.

Newman, Mark (2010). *Networks: An Introduction.* OUP Oxford, p. 784. ISBN: 0191500704.

Neyman, J. and E. S. Pearson (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 231.694-706, pp. 289–337. ISSN: 1364-503X. DOI: 10.1098/rsta.1933.0009.

Ng, Andrew Y., Michael I. Jordan, and Yair Weiss (2001). "On Spectral Clustering: Analysis and an algorithm". In: *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS.* MIT Press, pp. 849–856.

Nishikawa, T., A. E. Motter, Y.-C. Lai, and F. C. Hoppensteadt (2003). "Heterogeneity in Oscillator Networks: Are Smaller Worlds Easier to Synchronize?" In: *Phys. Rev. Lett.* 91, p. 014101.

Nolan, David S., Eric D. Rappin, and Kerry A. Emanuel (2007). "Tropical cyclogenesis sensitivity to environmental parameters in radiative-convective equilibrium". In: *Quarterly Journal of the Royal Meteorological Society* 133.629 B, pp. 2085–2107. ISSN: 00359009. DOI: 10.1002/qj.170.

Osborne, Martin J. and Ariel Rubinstein (1994). *A Course in Game Theory.* MIT Press, p. 352. ISBN: 0262650401.

Papalexiou, Simon Michael and Demetris Koutsoyiannis (2013). "Battle of extreme value distributions: A global survey on extreme daily rainfall". In: *Water Resources Research* 49.1, pp. 187–201. ISSN: 00431397. DOI: 10.1029/2012WR012557.

Peixoto, Tiago P. (2012). "Entropy of stochastic blockmodel ensembles". In: *Physical Review E* 85.5, p. 056122. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.85.056122.

– (2013). "Parsimonious Module Inference in Large Networks". In: *Physical Review Letters* 110.14, p. 148701. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.110.148701.

– (2014). "Hierarchical Block Structures and High-Resolution Model Selection in Large Networks". In: *Physical Review X* 4.1, p. 011047. ISSN: 2160-3308. DOI: 10.1103/PhysRevX.4.011047.

Peters, Ole and J. David Neelin (2006). "Critical phenomena in atmospheric precipitation". In: *Nature Physics* 2.June, p. 5. ISSN: 1745-2473. DOI: 10.1038/nphys314. arXiv: 0606076 [cond-mat].

– (2009). "Atmospheric Convection As a Continuous Phase Transition: Further Evidence". In: *International Journal of Modern Physics B* 23.28 & 29, pp. 5453–5465. ISSN: 0217-9792. DOI: 10.1142/S0217979209063778.

Peters, Ole, Christopher Hertlein, and Kim Christensen (2002). "A Complexity View of Rainfall". In: *Physical Review Letters* 88.1, p. 018701. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.88.018701. arXiv: 0201468 [cond-mat].

Peters, Ole, A. Deluca, A. Corral, J. D. Neelin, and C. E. Holloway (2010). "Universality of rain event size distributions". In: *Journal of Statistical Mechanics: Theory and Experiment* 11030, p. 16. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2010/11/P11030. arXiv: 1010.4201.

Peters, Ole, Kim Christensen, and J. David Neelin (2012). “Rainfall and Dragon-Kings”. In: *European Physical Journal: Special Topics* 205.1, pp. 147–158. ISSN: 19516355. DOI: 10.1140/epjst/e2012-01567-5.

Pikovsky, A., M. Rosenblum, and J. Kurths (2003). *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge Nonlinear Science Series. Cambridge University Press. ISBN: 9780521533522.

Pinto, R. D., P. Varona, A. R. Volkovski, A. Szücs, H. D. I. Abarbanel, and M. I. Rabinovich (2000). “Synchronous behavior of two coupled electronic neurons”. In: *Physiol Rev E* 62, pp. 2644–2656.

PyData Ecosystem. http://pydata.org/downloads/. Accessed: 2015-08-29.

Rao, A. Ramachandra, Rabindranath Jana, and Suraj Bandyopadhyay (1996). “a Chain Monte Carlo Method for Generating Random (0, 1)-Matrices with Given Marginals”. English. In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 58.2, pp. 225–242. ISSN: 0581572X.

Raymond, David J. and Sharon L. Sessions (2007). “Evolution of convection during tropical cyclogenesis”. In: *Geophysical Research Letters* 34.6. ISSN: 00948276. DOI: 10.1029/2006GL028607.

Sakaguchi, Hidetsugu (1988). “Cooperative Phenomena in Coupled Oscillator Systems under External Fields”. In: *Progress of Theoretical Physics* 79.1, pp. 39–46. DOI: 10.1143/PTP.79.39. eprint: http://ptp.oxfordjournals.org/content/79/1/39.full.pdf+html.

Santiago, A and RM Benito (2008). “An extended formalism for preferential attachment in heterogeneous complex networks”. In: *EPL (Europhysics Letters)* 82.5, p. 58004.

Scheel, M. L. M., M. Rohrer, Ch. Huggel, D. Santos Villar, E. Silvestre, and G. J. Huffman (2011). “Evaluation of TRMM Multi-satellite Precipitation Analysis (TMPA) performance in the Central Andes region and its dependency on spatial and temporal resolution”. In: *Hydrology and Earth System Sciences* 15.8, pp. 2649–2663. ISSN: 1607-7938. DOI: 10.5194/hess-15-2649-2011.

Serinaldi, Francesco and Chris G Kilsby (2014). “Rainfall extremes: Toward reconciliation after the battle of distributions”. In: *Water Resources Research* 50.1, pp. 336–352. ISSN: 0043-1397. DOI: 10.1002/2013WR014211.

Siqueira, JR and WB Rossow (2005). “Structural Characteristics of Convective Systems over South America Related to Cold-Frontal Incursions”. In: *Monthly Weather Review* 133.5, pp. 1045–1064. ISSN: 0027-0644. DOI: 10.1175/MWR2888.1.

Sokal, R.R., C.D. Michener, and University of Kansas (1958). *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas science bulletin. University of Kansas.

Sonnenschein, Bernard and Lutz Schimansky-Geier (2013). “Approximate solution to the stochastic Kuramoto model”. In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 88. ISSN: 15393755. DOI: 10.1103/PhysRevE.88.052111. arXiv: arXiv:1308.5629v1.

Stam, C.J. (2005). "Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field". In: *Clinical Neurophysiology* 116.10, pp. 2266 –2301. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2005.06.011.

Stolbova, V., P. Martin, B. Bookhagen, N. Marwan, and J. Kurths (2014). "Topology and seasonal evolution of the network of extreme precipitation over the Indian subcontinent and Sri Lanka". In: *Nonlinear Processes in Geophysics* 21.4, pp. 901–917. DOI: 10.5194/npg-21-901-2014.

Strehl, Alexander and Joydeep Ghosh (2003). "Cluster Ensembles - a Knowledge Reuse Framework for Combining Multiple Partitions". In: *J. Mach. Learn. Res.* 3, pp. 583–617. ISSN: 1532-4435. DOI: 10.1162/153244303321897735.

Strogatz, S.H. (1994). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering.* Advanced book program. Westview Press. ISBN: 9780738204536.

Thiel, M., J. Kurths, M.C. Romano, G. Károlyi, and A. Moura (2010). *Nonlinear Dynamics and Chaos: Advances and Perspectives.* Understanding Complex Systems. Springer Berlin Heidelberg. ISBN: 9783642046292.

Traxl, Dominik, Niklas Boers, and Jürgen Kurths (2016). "Deep Graphs - a general framework to represent and analyze heterogeneous complex systems across scales". In: *Chaos.* arXiv: 1604.00971.

Tulich, Stefan N. and George N. Kiladis (2012). "Squall Lines and Convectively Coupled Gravity Waves in the Tropics: Why Do Most Cloud Systems Propagate Westward?" In: *Journal of the Atmospheric Sciences* 69.10, pp. 2995–3012. ISSN: 0022-4928. DOI: 10.1175/JAS-D-11-0297.1.

Vapnik, V.N. (1998). *Statistical learning theory.* Adaptive and learning systems for signal processing, communications, and control. Wiley. ISBN: 9780471030034.

Varshney, Lav R., Beth L. Chen, Eric Paniagua, David H. Hall, and Dmitri B. Chklovskii (2011). "Structural Properties of the Caenorhabditis elegans Neuronal Network". In: *PLoS Comput Biol* 7.2, e1001066+. DOI: 10.1371/journal.pcbi.1001066.

Vera, C., W. Higgins, J. Amador, T. Ambrizzi, R. D. Garreaud, D. Gochis, D. Gutzler, D Lettenmaier, J. A. Marengo, C. R. Mechoso, J. Nogues-Paegle, P.L. Silva Dias, and C. Zhang (2006). "Toward a unified view of the American monsoon systems". In: *Journal of Climate* 19.20, pp. 4977–5000.

Virkar, Yogesh and Aaron Clauset (2014). "Power-law distributions in binned empirical data". In: *The Annals of Applied Statistics* 8.1, pp. 89–119. ISSN: 1932-6157. DOI: 10.1214/13-AOAS710. arXiv: arXiv:1208.3524v2.

Welsh, D.J.A. (2010). *Matroid Theory.* Dover books on mathematics. Dover Publications. ISBN: 9780486474397.

Whitaker, J.S. and C.A. Davis (1994). *Cyclogenesis in a saturated environment.* DOI: 10.1175/1520-0469(1994)051<0889:CIASE>2.0.CO;2.

Wiedermann, M., J. F. Donges, J. Heitzig, and J. Kurths (2013). "Node-weighted interacting network measures improve the representation of real-world complex systems". In: *EPL (Europhysics Letters)* 102.2, p. 28007.

Williams, Mark and Robert A. Houze Jr. (1987). *Satellite-Observed Characteristics of Winter Monsoon Cloud Clusters.* DOI: 10.1175/1520-0493(1987)115<0505:SOCOWM>2.0.CO;2.

Xu, Limei, Zhi Chen, Kun Hu, H. Eugene Stanley, and Plamen Ch. Ivanov (2006). “Spurious detection of phase synchronization in coupled nonlinear oscillators”. In: *Phys. Rev. E* 73 (6), p. 065201. DOI: 10.1103/PhysRevE.73.065201.

Zhou, Changsong, Lucia Zemanová, Gorka Zamora, Claus C. Hilgetag, and Jürgen Kurths (2006). “Hierarchical Organization Unveiled by Functional Connectivity in Complex Brain Networks”. In: *Phys. Rev. Lett.* 97 (23), p. 238103. DOI: 10.1103/PhysRevLett.97.238103.

Zipser, Edward J., Daniel J. Cecil, Chuntao Liu, Stephen W. Nesbitt, and David P. Yorty (2006). “Where Are the Most Intense Thunderstorms on Earth?” In: *Bulletin of the American Meteorological Society* 87.8, pp. 1057–1071. ISSN: 0003-0007. DOI: 10.1175/BAMS-87-8-1057.

Zulkafli, Zed, Wouter Buytaert, Christian Onof, Bastian Manz, Elena Tarnavsky, Waldo Lavado, and Jean-Loup Guyot (2014). “A comparative performance analysis of TRMM 3B42 (TMPA) versions 6 and 7 for hydrological applications over Andean-Amazon river basins”. In: *Journal of Hydrometeorology* 15.2, pp. 581–592. ISSN: 1525-755X. DOI: 10.1175/JHM-D-13-094.1.

Selbständigkeitserklärung

Ich erkläre, dass ich die Dissertation selbständig und nur unter Verwendung der von mir gemäß §7 Abs. 3 der Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt-Universität zu Berlin Nr. 126/2014 am 18.11.2014 angegebenen Hilfsmittel angefertigt habe.

Berlin, den 16. Mai 2017 Dominik Traxl