



University of Pennsylvania  
**ScholarlyCommons**

---

Publicly Accessible Penn Dissertations


---

2022

## The Network Science Of Distributed Representational Systems

Harang Ju  
*University of Pennsylvania*

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Neuroscience and Neurobiology Commons](#)

---

### Recommended Citation

Ju, Harang, "The Network Science Of Distributed Representational Systems" (2022). *Publicly Accessible Penn Dissertations*. 5159.  
<https://repository.upenn.edu/edissertations/5159>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/5159>  
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).

---

# The Network Science Of Distributed Representational Systems

## Abstract

From brains to science itself, distributed representational systems store and process information about the world. In brains, complex cognitive functions emerge from the collective activity of billions of neurons, and in science, new knowledge is discovered by building on previous discoveries. In both systems, many small individual units—neurons and scientific concepts—interact to inform complex behaviors in the systems they comprise. The patterns in the interactions between units are telling; pairwise interactions not only trivially affect pairs of units, but they also form structural and dynamic patterns with more than just pairs, on a larger scale of the network. Recently, network science adapted methods from graph theory, statistical mechanics, information theory, algebraic topology, and dynamical systems theory to study such complex systems. In this dissertation, we use such cutting-edge methods in network science to study complex distributed representational systems in two domains: cascading neural networks in the domain of neuroscience and concept networks in the domain of science of science.

In the domain of neuroscience, the brain is a system that supports complex behavior by storing and processing information from the environment on long time scales. Underlying such behavior is a network of millions of interacting neurons. Many recent studies measure neural activity on the scale of the whole brain with brain regions as units or on the scale of brain regions with individual neurons as units. While many studies have explored the neural correlates of behaviors on these scales, it is less explored how neural activity can be decomposed into low-level patterns. Network science has shown potential to advance our understanding of large-scale brain networks, and here, we apply network science to further our understanding of low-level patterns in small-scale neural networks. Specifically, we explore how the structure and dynamics of biological neural networks support information storage and computation in spontaneous neural activity in slice recordings of rodent brains. Our results illustrate the relationships between network structure, dynamics, and information processing in neural systems.

In the domain of science of science, the practice of science itself is a system that discovers and curates information about the physical and social world. For centuries, philosophers, historians, and sociologists of science have theorized about the process and practice of scientific discovery. Recently, the field of science of science has emerged to use a more data-driven approach to quantify the process of science. However, it remains unclear how recent advances in science of science either support or refute the various theories from the philosophies of science. Here, we use a network science approach to operationalize theories from prominent philosophers of science, and we test those theories using networks of hyperlinked articles in Wikipedia, the largest online encyclopedia. Our results support a nuanced view of philosophies of science—that science does not grow outward, as many may intuit, but by filling in gaps in knowledge.

In this dissertation, we examine cascading neural networks first in Chapters 2 through 4 and then concept networks in Chapter 5. The studies in Chapters 2 to 4 highlight the role of patterns in the connections of neural networks in storing information and performing computations. The study in Chapter 5 describes patterns in the historical growth of concept networks of scientific knowledge from Wikipedia. Together, these analyses aim to shed light on the network science of distributed representational systems that store and process information about the world.

## Degree Type

Dissertation

## Degree Name

Doctor of Philosophy (PhD)

---

**Graduate Group**

Neuroscience

**First Advisor**

Dani S. Bassett

**Subject Categories**

Neuroscience and Neurobiology

THE NETWORK SCIENCE OF DISTRIBUTED REPRESENTATIONAL SYSTEMS

Harang Ju

A DISSERTATION

in

Neuroscience

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Dani S. Bassett, J. Peter Skirkanich Professor of Bioengineering, Electrical & Systems Engineering, Physics & Astronomy, Neurology, & Psychiatry

Graduate Group Chairperson

Joshua I. Gold, Professor of Neuroscience

Dissertation Committee

Maria Neimark Geffen, Associate Professor of Otorhinolaryngology, Neuroscience & Neurology

Eleni Katifori, Associate Professor of Physics & Astronomy

Erol Akçay, Associate Professor of Biology



THE NETWORK SCIENCE OF DISTRIBUTED REPRESENTATIONAL SYSTEMS  
COPYRIGHT

2022

This work is licensed under the  
Creative Commons  
CC BY 4.0  
License

To view a copy of this license, visit

<https://creativecommons.org/licenses/by/4.0/>

*This dissertation is dedicated to my parents for being a constant source of love, support, and foresight and to my younger sister whose discipline and fearlessness inspire me daily.*

## ACKNOWLEDGEMENT

First and foremost, I would like to thank my thesis supervisor, Dani Bassett, who is as kind as they are brilliant. A scientific mentor sees you through your best and your worst, when you make a new discovery or when you try something for months on end without any results. Throughout all these times, Dani nurtured me through kindness, patience, and curiosity. I could not have asked for a better mentor, and for that I am so thankful.

I am also immensely grateful for the opportunity to have worked with the generous and brilliant people in Dani's lab. My lab mates are my academic family, and I learned so much from them. In the Bassett lab, I could walk over and banter with anyone in the lab, and—in that brief moment—we would transcend the individual to explore the unknown.

I am grateful to all of the faculty and staff at Penn who have provided me with mentorship and support. I want to thank Drs. Maria Geffen, Eleni Katifori, and Erol Akçay for serving on my thesis committee. All of their feedback, advice, and oversight was invaluable to my development as a scientist. I am extremely grateful to Dr. Josh Gold and Christine Clay, the director and the former coordinator of the Neuroscience Graduate Group, respectively, who kindly and passionately supported me in my formal education in neuroscience.

I am deeply grateful for my family and friends. I want to thank my parents for loving me and pushing me to be the best that I can be. I want to thank my little sister, YeRang, for always being there to listen to my ramblings. I am grateful for my girlfriend, Hailey Park, for being the salt of my life, without whom my Ph.D. years would have been but bland.

Finally, I want to thank my God, Christ Jesus, for loving me unconditionally and for creating a beautiful world for us to explore.

# ABSTRACT

## THE NETWORK SCIENCE OF DISTRIBUTED REPRESENTATIONAL SYSTEMS

Harang Ju

Dani S. Bassett

From brains to science itself, distributed representational systems store and process information about the world. In brains, complex cognitive functions emerge from the collective activity of billions of neurons, and in science, new knowledge is discovered by building on previous discoveries. In both systems, many small individual units—neurons and scientific concepts—interact to inform complex behaviors in the systems they comprise. The patterns in the interactions between units are telling; pairwise interactions not only trivially affect pairs of units, but they also form structural and dynamic patterns with more than just pairs, on a larger scale of the network. Recently, network science adapted methods from graph theory, statistical mechanics, information theory, algebraic topology, and dynamical systems theory to study such complex systems. In this dissertation, we use such cutting-edge methods in network science to study complex distributed representational systems in two domains: cascading neural networks in the domain of neuroscience and concept networks in the domain of science of science.

In the domain of neuroscience, the brain is a system that supports complex behavior by storing and processing information from the environment on long time scales. Underlying such behavior is a network of millions of interacting neurons. Many recent studies measure neural activity on the scale of the whole brain with brain regions as units or on the scale of brain regions with individual neurons as units. While many studies have explored the neural correlates of behaviors on these scales, it is less explored how neural activity can be decomposed into low-level patterns. Network science has shown potential to advance our understanding of large-scale brain networks, and here, we apply network science to further our understanding of low-level patterns in small-scale neural networks. Specifically, we explore

how the structure and dynamics of biological neural networks support information storage and computation in spontaneous neural activity in slice recordings of rodent brains. Our results illustrate the relationships between network structure, dynamics, and information processing in neural systems.

In the domain of science of science, the practice of science itself is a system that discovers and curates information about the physical and social world. For centuries, philosophers, historians, and sociologists of science have theorized about the process and practice of scientific discovery. Recently, the field of science of science has emerged to use a more data-driven approach to quantify the process of science. However, it remains unclear how recent advances in science of science either support or refute the various theories from the philosophies of science. Here, we use a network science approach to operationalize theories from prominent philosophers of science, and we test those theories using networks of hyperlinked articles in Wikipedia, the largest online encyclopedia. Our results support a nuanced view of philosophies of science—that science does not grow outward, as many may intuit, but by filling in gaps in knowledge.

In this dissertation, we examine cascading neural networks first in Chapters 2 through 4 and then concept networks in Chapter 5. The studies in Chapters 2 to 4 highlight the role of patterns in the connections of neural networks in storing information and performing computations. The study in Chapter 5 describes patterns in the historical growth of concept networks of scientific knowledge from Wikipedia. Together, these analyses aim to shed light on the network science of distributed representational systems that store and process information about the world.

## TABLE OF CONTENTS

ACKNOWLEDGEMENT . . . . .	iv
ABSTRACT . . . . .	v
LIST OF TABLES . . . . .	viii
LIST OF ILLUSTRATIONS . . . . .	ix
CHAPTER 1 : GENERAL INTRODUCTION . . . . .	1
CHAPTER 2 : DYNAMIC REPRESENTATIONS IN NEURAL NETWORKS . . . . .	22
CHAPTER 3 : MEMORY IN CASCADING NEURAL NETWORKS . . . . .	56
CHAPTER 4 : LOGIC IN CASCADING NEURAL NETWORKS . . . . .	117
CHAPTER 5 : GROWTH IN CONCEPT NETWORKS . . . . .	141
CHAPTER 6 : GENERAL CONCLUSIONS . . . . .	203

## LIST OF TABLES

TABLE 3.1	Network parameters for all simulations . . . . .	112
TABLE 5.1	Network measures . . . . .	198
TABLE 5.2	Network metrics for subjects . . . . .	199
TABLE 5.3	Core-periphery lead-lag $t$ -tests for all subjects . . . . .	200

## LIST OF ILLUSTRATIONS

FIGURE 2.1	Neural representations and tools to analyze them . . . . .	25
FIGURE 2.2	Network models abstract neural systems . . . . .	28
FIGURE 2.3	Integrating network models and neural representations . . . . .	30
FIGURE 2.4	Dynamic representations in networked neural systems . . . . .	33
FIGURE 2.5	Dynamic representations as trajectories in neural state space . . . .	43
FIGURE 3.1	A linear dynamical system estimates spiking in a stochastic model .	59
FIGURE 3.2	Network topology constrains cascade duration . . . . .	63
FIGURE 3.3	Cycles and strong connections facilitate long cascades . . . . .	65
FIGURE 3.4	Network controllability is tightly linked with cascade duration . . .	69
FIGURE 3.5	A stimulus well-recovered when it generates long-lasting cascades . .	74
FIGURE 3.6	Numerical validation of Markov formulation . . . . .	99
FIGURE 3.7	Linear dynamical model predicts average firing rate . . . . .	101
FIGURE 3.8	Strong edge weights in cycles produce long cascades . . . . .	102
FIGURE 3.9	Finite average controllability is correlated with mean cascade duration	103
FIGURE 3.10	Longer cascade duration allows stimulus recovery . . . . .	104
FIGURE 3.11	Collisions may occur frequently in living neuronal systems . . . . .	105
FIGURE 3.12	Controllability constrains cascade duration . . . . .	106
FIGURE 3.13	Reverberations in cascading dynamics . . . . .	107
FIGURE 3.14	Likelihood ratio test between models . . . . .	108
FIGURE 3.15	Exponent relations for criticality in MEA recordings . . . . .	110
FIGURE 3.16	Mean absolute errors of VAR models from 25 recordings . . . . .	111
FIGURE 4.1	Neural logic gates and how to find them . . . . .	119
FIGURE 4.2	Model of logic gates . . . . .	121
FIGURE 4.3	Logical gates in cortical neurons . . . . .	123
FIGURE 4.4	Computations in hippocampal neurons . . . . .	124
FIGURE 4.5	<i>In vitro</i> neural development . . . . .	125
FIGURE 4.6	Cluster transition probability . . . . .	126
FIGURE 4.7	All clusters of the “XOR” model . . . . .	136
FIGURE 4.8	Participation of neurons in clusters . . . . .	137
FIGURE 4.9	Cascade statistics . . . . .	138
FIGURE 4.10	Neural logic gates require spiketimes with millisecond resolution . .	139
FIGURE 4.11	Neural logic gates require cascades . . . . .	140
FIGURE 5.1	Building a growing concept network from Wikipedia . . . . .	144
FIGURE 5.2	Real concept networks maintain shorter and fewer gaps . . . . .	147
FIGURE 5.3	Concept networks undergo a signature pattern in structural stability	150



FIGURE 5.4	Concept networks undergo a signature pattern in structural stability	151
FIGURE 5.5	Core-periphery lead-lag relationships . . . . .	188
FIGURE 5.6	Some core nodes are born early while others are not. . . . .	189
FIGURE 5.7	Statistics of tf-idf vectors in an example real network . . . . .	190
FIGURE 5.8	Paradigm shift signature breaks in edge-rewired networks . . . . .	191
FIGURE 5.9	Degree distributions of edge-rewired null networks . . . . .	191
FIGURE 5.10	The structural stability signature robust to interslice weight . . . .	192
FIGURE 5.11	Changepoint detection using PELT versus binary segmentation . . .	193
FIGURE 5.12	Shorter time horizons for impulse response and cavity participation	194
FIGURE 5.13	Robustness to changes in the year of nodes . . . . .	195
FIGURE 5.14	Visualizations of four example networks . . . . .	196
FIGURE 5.15	Coreness of thresholded networks . . . . .	197

# CHAPTER 1

## GENERAL INTRODUCTION

### 1.1. Network science and distributed representational systems

Network science uses theories and methods from graph theory, statistical mechanics, information theory, algebraic topology, and dynamical systems theory to study complex systems of interacting units. In their key paper, Watts et al. (1998) demonstrated that networks representations of many complex systems have small-world properties, where connection patterns lie between full order and full disorder. Dynamical systems that can be represented as small world networks exhibit enhanced signal-propagation speed, computational power, and synchronizability. Importantly, they used (i) graph theory to characterize a spectrum of network structures, from ordered to random, (ii) data to identify “small world” networks in real systems, and (iii) dynamical systems to model system performance as a function of network structure. Since then, studies have quantified the network structures of complex systems and how they relate to system performance across scholarly fields, from cell biology and neuroscience to statistical mechanics and science of science, (Newman, 2001; Albert et al., 2002; Barabási et al., 2004; Bassett and Bullmore, 2006; Kwak et al., 2010; Bassett and Sporns, 2017).

In this dissertation, we will use the network science approach to study complex systems that represent information about the environment among distributed units, which we label as *distributed representational systems*, in two distinct domains: cascading neural networks and scientific knowledge databases. First, biological neural networks represent features of the environment among distributed units called neurons. The networks display cascades, which are spontaneous activity that propagate across a network of synaptic connections between neurons (Beggs and Plenz, 2003) and are poised at a regime that facilitates the optimal storage and transmission of information across the network (Beggs, 2004; Halderman et al., 2005; Larremore et al., 2011). Moreover, such networks adapt to maintain such

optimal information processing properties (Shew, Clawson, et al., 2015). Second, we explore Wikipedia, which is the largest online encyclopedia and a database of scientific knowledge. Such databases—instead of using networks of dynamically interacting units—store information in units of articles, where related articles are hyperlinked to form a network of concepts. Here, we will explore these diverse distributed representational systems from the perspective of network science to uncover how the network structure of a system relates to its ability to store and process information.

## 1.2. Network neuroscience

The “research programme” of neuroscience builds on the core concept of the *neuron doctrine*, which states that brain function is based on organized neural networks (Llinás, 2003). Neural networks receive and process stimuli from the environment to inform cognition and behavior, such as the visual recognition of faces (Adolphs, 2003). Thus, a central question in neuroscience is how connections between neurons support cognition and behavior for an organism to identify and act upon environmental stimuli. To identify environmental stimuli, an organism must form representations of its environment in its brain, and the scope of representations has been experimentally demonstrated across a wide spectrum of cognitive functions, from mapping social hierarchies to predicting future events (Summerfield et al., 2006; Tavares et al., 2015).

To aid in the study of how organisms store, transmit, and process information, network models distill the complexity of the brain into its units of brain regions or neurons and their interactions across white matter connections or synapses (Bassett, Zurn, et al., 2018). By formulating a brain or a neuronal population as a network, network neuroscience can take advantage of theories and methods from a range of disciplines, from graph theory and statistics to physics and engineering (Watts et al., 1998; Bassett and Sporns, 2017; Lynn et al., 2019). For example, networks are often formulated as a connectivity matrix  $A$ , where element  $a_{ij}$  of the matrix  $A$  scales the interaction between units  $i$  and  $j$ .

In neuroscience, networks have been used to model neural systems primarily at two spatial scales: a larger scale of whole brains with regions as units, and a smaller scale of neurons as units. At the larger scale, studies have used data from tractography of white matter connections or BOLD signals that are correlated with regional brain activity (Friston, 2011; Hermundstad et al., 2013). Such network models have been used to understand the neural correlates of cognition and also to seek potential clinical uses (Medaglia et al., 2018; Jeganathan et al., 2018; Cornblath et al., 2020; Stiso et al., 2019). On the smaller, cellular scale, neural connections and their underlying computational function have often been inferred through neural dynamics (Hopfield, 1982; Ben-Yishai et al., 1995; Wang, 2002).

Using advanced techniques in network neuroscience, recent studies have characterized the network structures of brains at a range of scales. Large-scale brain networks have been shown to be small-world with long distance white matter connections and modular with many anatomically and structurally distinct brain regions (Bertolero et al., 2015; Bassett and Bullmore, 2017). On the scale of neuronal populations, neural networks are also modular with hubs and clusters (Shimono et al., 2014). Furthermore, cortical neurons have patterns of connectivity that are common throughout the cortex, including bidirectionally connected neurons and higher-order network motifs (Wang, Markram, et al., 2006; Lefort et al., 2009; Ko et al., 2011; Markram, 1997; Song et al., 2005; Perin et al., 2011).

The structure of brain networks has been shown to be critical to brain dynamics that underlie information processing. One of the simplest models of brain activity is that of a linear system (Kailath, 1980; Becker et al., 2018; Nozari et al., 2021). A linear system assumes that vectors of neural activity  $x$  evolve according to pairwise connections between units, represented in a connectivity matrix  $A$ , as  $x(t+1) = Ax(t) + Bu(t)$ , where  $B$  scales inputs  $u(t)$  to the system. A linear systems model provides closed form solutions to many properties of the system, including controllability and stability (Kailath, 1980; Pasqualetti et al., 2014). Using these tools, recent studies have explored how the brain may react to control by other brain regions or exogenous stimuli (Muldoon et al., 2016; Gu et al., 2017; Tang et al., 2018; Medaglia

et al., 2018).

While dynamical systems theory relates network structure to dynamics, advanced techniques in information theory and statistics have revealed relationships between network dynamics and information processing. Network models that use transfer entropy to identify synaptic connections have revealed common patterns in the network structures of cortical neurons, including clusters, rich clubs, and communities (Ito et al., 2011; Shimono et al., 2014; Nigam et al., 2016; Timme et al., 2016). Moreover, certain structural patterns, such as hubs and rich clubs, have been found to perform most of the computation in populations of cortical neurons (Chen et al., 2010; Faber et al., 2019). While many studies aim to determine how the brain’s network structure may support information processing and thus cognition, it remains challenging to directly measure the relationships between neural dynamics, connectivity, and computation (Watts et al., 1998; Honey et al., 2007; Eliasmith et al., 2012).

### 1.3. Neuronal avalanches

One interesting phenomenon in which neural dynamics are tightly coupled to network structure is that of neuronal avalanches, which we study in this dissertation. Neuronal avalanches, first discovered in spontaneously active neurons, consist of bursts of neuronal activity, whose spatial and temporal correlations are scale-free (Beggs and Plenz, 2003). Importantly, neuronal avalanches were shown to operate in a regime with optimal information processing properties, including optimal information transmission (Beggs and Plenz, 2003; Shew, Yang, Yu, et al., 2011), information storage (Haldeman et al., 2005), computational power (Bertschinger et al., 2004), and dynamic range (Kinouchi et al., 2006; Shew, Yang, Petermann, et al., 2009; Larremore et al., 2011). The term *avalanche* is borrowed from the field of statistical mechanics in which Bak et al. (1987) demonstrated *self-organizing criticality* in an Abelian sandpile model. In the model, grains of sand are dropped until they form a pile of sand; dropping a grain on the pile triggers an avalanche that ripples down the pile. Interestingly, the statistical methanics of avalanches were shown to have

scale-free spatial and temporal correlations such that there is no “scale”—or mathematically, an expected value—for the duration or size (the number of units in an event) of avalanches. Thus, distributions of size and duration of avalanches are heavy-tailed and form power laws.

Critical systems display statistics beyond power-law distributions of avalanche size and duration, which can be displayed by non-critical systems. An important additional test for criticality is the exponent relation, which states that size and duration of events should scale proportionally (Friedman et al., 2012). Mathematically, the power law exponents,  $\alpha$  and  $\tau$ , of the distributions of event duration  $d$  and size  $s$ , respectively, are related to the power law exponent of cascade size given duration (Friedman et al., 2012). Recent studies use the exponent relation to identify criticality in their experimental systems (Shew, Clawson, et al., 2015; Ponce-Alvarez et al., 2018; Fontenele et al., 2019). However, many studies have found that not all systems that have power law distributions in size and duration are at criticality, such as molecular chaos models (Touboul et al., 2017). Moreover, subsampling of systems often leads to a strong overestimation of the stability in neural systems, which should can be corrected using multistep regression estimation (Wilting et al., 2018).

Neuronal avalanches have been demonstrated to exist in the brain across a range of experimental measurements. The experimental evidence of neuronal avalanches spans a range of methods *in vitro* (Beggs and Plenz, 2003; Beggs, 2004), *in vivo* (Gireesh et al., 2008; Petermann et al., 2009; Hahn et al., 2010; Shriki et al., 2013; Bellay et al., 2015; Ponce-Alvarez et al., 2018), and *ex vivo* (Shew, Clawson, et al., 2015) in a variety of organisms, including humans. However, the hypothesis of criticality in the brain remains somewhat controversial (Beggs and Timme, 2012; Wilting et al., 2019); some neural systems are slightly sub-critical with “reverberating” dynamics, which has been observed *in vivo* from spike recordings of cat, monkey, and rats (Wilting et al., 2018). Because a neural system often receives stimuli from the environment that push the system out of the regime of criticality, neurons may often reside out of criticality but quickly adapt to near criticality (Shew, Clawson, et al., 2015).

#### 1.4. Neuronal cascades

The near-criticality of many neural systems begs the question: what kinds of information processing do such systems perform? Specifically, how do the bursts of activity perform computations beyond those of storing and relaying information? How does neural activity synthesize and transform inputs into outputs? To set briefly aside the debate regarding criticality in the brain, we began to use the term *cascades* versus *avalanches* to discuss spontaneous activity without the strict requirements for criticality (Ju, Kim, et al., 2020). A “cascade” refers to a burst of neural activity, whose continuous activity is presumed to be due to synaptic interactions between neurons. Indeed, some recordings of neural activity were shown to be slightly below critical in the analyses by other studies (Wilting et al., 2018) and in our analyses (Ju, Kim, et al., 2020). The distributions of such sub-critical systems can be modeled using an exponentially truncated power law  $p(x) \sim x^{-\alpha} e^{-x/\tau}$  where  $\tau$  is a constant that modulates the exponential truncation (Denisov et al., 2016; Murphy et al., 2019).

The network science perspective begets inquiry into the structure of networks underlying these complex dynamics. While neural cascades at coarse time resolutions of seconds often have the appearance of a single burst of activity, cascades at finer time resolutions of milliseconds are indeed richly varied yet stable spatiotemporal patterns of activity (Beggs, 2004). Additionally, a separate line of research observed motifs in the patterns in synaptic connectivity in cortical neurons. For example, cortical neurons are often strongly and bidirectionally connected to each other (Wang, Markram, et al., 2006; Lefort et al., 2009; Ko et al., 2011) and form even higher-order motifs in clusters of neurons (Markram, 1997; Song et al., 2005; Perin et al., 2011). While these motifs can theoretically support short-term information storage (Rodriguez et al., 2001; Fiete et al., 2010; Daie et al., 2015; Brunel, 2016), the relationship between neural cascades and the information processing properties remains unclear.

While the avalanche literature views neuronal information processing from a statistical mechanics perspective—thus quantifying information processing of a system as a whole—recent studies use state-of-the-art information theory to begin to quantify computations beyond information storage and transmission in neural populations (Shimono et al., 2014; Nigam et al., 2016; Faber et al., 2019; Lynn et al., 2019; Ju and Bassett, 2020). Specifically, partial information decomposition was recently developed as a way to decompose information processing from multiple sources that are independent, dependent, and synergistic (Wibral, Priesemann, et al., 2017; Wibral, Finn, et al., 2017). Synergistic information processing takes many inputs and maps them to outputs such that all inputs are required to determine the output. These new advances reveal a new frontier in the study of how near-critical neural systems may perform complex computations (Bertschinger et al., 2004).

## 1.5. Science of science

In addition to cascading neural networks, scientific knowledge itself is the second information system that we will explore in this dissertation.

What is the same? What is different? How are they both distributed representational systems? Do they differ in their representations? Do they differ in the manner in which those representations are distributed? Why study the two together? What do you gain that you wouldn't have acquired if you had studied either alone. Need some narrative of integration.

To motivate our study of the network science of scientific progress, we first visit the idea of *multiple discoveries*. In 1974, Merton observed the phenomenon, or hypothesis, of *multiple discoveries* which states that for many scientific discoveries, numerous groups or individuals make the same discovery independently and contemporaneously (Merton, 1974). Many famous discoveries were *multiple discoveries*. Calculus was discovered independently and contemporaneously by Sir Isaac Newton and Gottfried Wilhelm Leibniz, and evolution was discovered by both Charles Darwin and Alfred Russel Wallace (Merton, 1974). The phe-



nomenon of multiple discoveries points to perhaps an intuitive “law” in scientific discovery: certain discoveries require that, to discover them, one must have access to certain other knowledge or materials. Newton famously professed a similar sentiment, “If I have seen further, it is by standing on the shoulders of giants” (Newton, 1675). Conversely, access to certain knowledge or materials may *facilitate* certain discoveries. So one may naturally ask whether there are patterns to how certain sets of knowledge may lead to certain discoveries.

To begin to study whether and how existing knowledge supports discoveries, we must review to prominent theories by philosophers of science. For centuries, scholars who have studied scientific progress have postulated that various processes underlie scientific progress. Popper, in his classic paper “Conjectures and refutations: the growth of scientific knowledge,” described scientific progress as a sequence of theories in which previous ones are falsified by new ideas (Popper, 1968). Then building on his work, Kuhn, Feyerabend, and Lakatos each offered three prominent and influential philosophies of science (Kuhn et al., 2012; Feyerabend, 2010; Lakatos, 1968). In 1962, Kuhn described scientific progress as periods of normal science, in which scientists “solved puzzles” within a paradigm, which is the set of basic concepts and experimental practices of a scientific discipline (Kuhn et al., 2012). Periods of normal science are separated from one another by paradigm shifts that overturn the paradigm. In 1970, Lakatos suggested an alternate view that scientific progress is based on “research programmes”, in which knowledge expands from a common core set of scientific ideas and experimental practices (Lakatos, 1968). In 1975, Feyerabend dismissed any single mechanism for scientific progress (Feyerabend, 2010).

Recently, the field of science of science has begun to explore questions in scientific progress in a more quantitative manner. New studies in science of science have sought to quantify and predict scientific research and the resulting outcomes from data-driven and complex systems perspectives (Wang, Song, et al., 2013; Clauset et al., 2017; Zeng et al., 2017; Fortunato et al., 2018). Other studies inquire into the institutional, personal, and societal conditions that do or do not support scientific discovery (Sinatra et al., 2016; Chemla et al., 2017; Helmer

et al., 2017; Astegiano et al., 2019; Wu et al., 2019; Nagaraj et al., 2020; Robinson-Garcia et al., 2020). Despite the recent advances in data-driven studies on scientific practice, it remains unclear whether the prominent philosophies of science are supported by currently available sources of data on the practice of science.

In our study in science of science, we aim to begin to bridge the gap between science of science and the prominent theories of the philosophers of science. To do so, we use methods from network science to formulate a scientific body of knowledge as a network of concepts and quantify patterns in the inter-concept relations (Siew et al., 2019). Recently, networks have been useful in studying the changes in the inter-relations between concepts, as “semantic networks” formed by the co-occurrences of words in books (Christianson et al., 2020). Moreover, concept networks have proven to be powerful tools for probing questions about the exploration of knowledge across Wikipedia articles (Lydon-Staley et al., 2021). We perform these analyses on hyperlinked articles from Wikipedia, the largest online encyclopedia. Taken together, recent developments in the study of science illustrate that the time is ripe to study patterns in scientific discovery itself through a quantitative lens.

## **1.6. Architecture of this thesis**

In this dissertation, the overarching goal is to shed light on relationships between the structure and function of networks in distributed representational systems. Such systems use a distributed network of units that together form representations of the environment. In biological neural networks, distributed neurons interact to represent and model their environment, and in science, distributed concepts refer to one another to represent and model the physical and social world. Here, we aim to explore how patterns of network structure relate to the function of neural networks in neuroscience and concept networks in science of science.

In Chapter 2, we review the literature on two distinct lines of research in neuroscience: network models and neural representations. While the former describes patterns of neural

interactions, the latter describes patterns of neural activity as correlates of representations of an organism’s environment. We then propose a framework of dynamic neural representations that unites the two fields to further our understanding of the neural models of an ever-changing environment.

In Chapter 3, we employ a dynamic network model based on spontaneous spiking activity from the slice recordings of mouse somatosensory cortex. We use network measures, linear systems theory, and information theory to study the relationship between network structure and memory capacity in neural systems. We identify network structures at both the microscopic and macroscopic scales, and demonstrated how they can support the retention of information across time.

In Chapter 4, we expand the study of neural interactions from pairwise to triplet-wise interactions by operationalizing neural “logic gates” as firing probabilities conditional on two other neurons. In spontaneously spiking neurons in slice recordings of mouse somatosensory cortex and rat hippocampus, we quantified non-trivial, triplet-wise interactions between neurons. Those interactions differed across brain regions and throughout the *in vitro* development of synaptic connections.

In Chapter 5, we operationalize and test theories from prominent philosophers of science using network measures, algebraic topology, and linear systems theory. By creating concept networks that “grow” throughout history from hyperlinked Wikipedia articles, we found that the body of knowledge does not grow outward, as many may intuit, but rather by filling in gaps in knowledge. Such gap-filling is also deemed important in scientific communities as concepts that either create or fill in knowledge gaps are more influential and more often awarded Nobel prizes.

Taken together, the studies demonstrate how the organization of systems with distributed networks of units can represent and model the environment. In neural systems, the distributed activity of neurons interact with one another to store, compute, and transmit

information from environmental stimuli, and in science, distributed concepts relate to one another to model the physical and social world. These two systems reveal two different ways of modeling the environment—first by distributed activity and second by distributed relations—both via a distributed network of units. These studies begin to pave the way to better understanding the low-level relationships between network structure and function that may eventually facilitate the engineering and control of these systems.

## REFERENCES

- Adolphs, Ralph (Mar. 1, 2003). “Cognitive neuroscience of human social behaviour.” In: *Nature Reviews Neuroscience* 4, 165 EP.
- Albert, Réka and Albert-László Barabási (Jan. 30, 2002). “Statistical mechanics of complex networks.” In: *Reviews of Modern Physics* 74.1, pp. 47–97. DOI: 10.1103/RevModPhys.74.47. (Visited on 05/14/2022).
- Astegiano, Julia, Esther Sebastián-González, and Camila de Toledo Castanho (June 2019). “Unravelling the gender productivity gap in science: a meta-analytical review.” In: *Royal Society Open Science* 6.6, p. 181566. DOI: 10.1098/rsos.181566.
- Bak, Per, Chao Tang, and Kurt Wiesenfeld (July 1987). “Self-organized criticality: An explanation of the  $1/f$  noise.” In: *Physical Review Letters* 59.4, pp. 381–384.
- Barabási, Albert-László and Zoltán N. Oltvai (Feb. 2004). “Network biology: understanding the cell’s functional organization.” In: *Nature Reviews Genetics* 5.2, pp. 101–113. DOI: 10.1038/nrg1272. (Visited on 05/14/2022).
- Bassett, Danielle S. and Edward T. Bullmore (Oct. 2017). “Small-World Brain Networks Revisited.” In: *The Neuroscientist* 23.5, pp. 499–516. DOI: 10.1177/1073858416667720.
- Bassett, Danielle S., Perry Zurn, and Joshua I. Gold (2018). “On the nature and use of models in network neuroscience.” In: *Nature Reviews Neuroscience* 19.9, pp. 566–578. DOI: 10.1038/s41583-018-0038-8.
- Bassett, Danielle S and Olaf Sporns (2017). “Network neuroscience.” In: *Nature Neuroscience* 20.3, pp. 353–364. DOI: 10.1038/nn.4502.
- Bassett, Danielle Smith and Ed Bullmore (2006). “Small-World Brain Networks.” In: *The Neuroscientist* 12.6, pp. 512–523. DOI: 10.1177/1073858406293182.
- Becker, Cassiano O, Danielle S Bassett, and Victor M Preciado (2018). “Large-scale dynamic modeling of task-fMRI signals via subspace system identification.” In: *Journal of Neural Engineering* 15.6, p. 066016. DOI: 10.1088/1741-2552/aad8c7.

- Beggs, J. M. (2004). “Neuronal Avalanches Are Diverse and Precise Activity Patterns That Are Stable for Many Hours in Cortical Slice Cultures.” In: *Journal of Neuroscience* 24.22, pp. 5216–5229. DOI: 10.1523/JNEUROSCI.0540-04.2004.
- Beggs, John M. and Dietmar Plenz (2003). “Neuronal Avalanches in Neocortical Circuits.” In: *Journal of Neuroscience* 23.35, pp. 11167–11177. DOI: 10.1523/JNEUROSCI.23-35-11167.2003.
- Beggs, John and Nicholas Timme (2012). “Being Critical of Criticality in the Brain.” In: *Frontiers in Physiology* 3, p. 163. DOI: 10.3389/fphys.2012.00163.
- Bellay, Timothy et al. (July 2015). “Irregular spiking of pyramidal neurons organizes as scale-invariant neuronal avalanches in the awake state.” In: *eLife* 4. Ed. by Frances K Skinner, e07224. DOI: 10.7554/eLife.07224.
- Ben-Yishai, R., R L Bar-Or, and H Sompolinsky (1995). “Theory of orientation tuning in visual cortex.” In: *Proceedings of the National Academy of Sciences* 92.9, pp. 3844–3848. DOI: 10.1073/pnas.92.9.3844.
- Bertolero, Maxwell A., B. T. Thomas Yeo, and Mark D’Esposito (2015). “The modular and integrative functional architecture of the human brain.” In: *Proceedings of the National Academy of Sciences* 112.49. DOI: 10.1073/pnas.1510619112.
- Bertschinger, Nils and Thomas Natschläger (2004). “Real-Time Computation at the Edge of Chaos in Recurrent Neural Networks.” In: *Neural Computation* 16.7, pp. 1413–1436. DOI: 10.1162/089976604323057443.
- Brunel, Nicolas (Apr. 11, 2016). “Is cortical connectivity optimized for storing information?” In: *Nature Neuroscience* 19, 749 EP.
- Chemla, Karine and Evelyn Fox Keller, eds. (2017). *Cultures without culturalism: the making of scientific knowledge*. Durham: Duke University Press. 410 pp.
- Chen, Wei et al. (Jan. 2010). “A few strong connections: optimizing information retention in neuronal avalanches.” In: *BMC neuroscience* 11, pp. 3–3. DOI: 10.1186/1471-2202-11-3.
- Christianson, Nicolas H., Ann Sizemore Blevins, and Danielle S. Bassett (July 2020). “Architecture and evolution of semantic networks in mathematics texts.” In: *Proceedings of the*

- Royal Society A: Mathematical, Physical and Engineering Sciences* 476.2239, p. 20190741.  
DOI: 10.1098/rspa.2019.0741.
- Clauset, Aaron, Daniel B. Larremore, and Roberta Sinatra (Feb. 3, 2017). “Data-driven predictions in the science of science.” In: *Science* 355.6324, pp. 477–480. DOI: 10.1126/science.aal4217.
- Cornblath, Eli J. et al. (2020). “Temporal sequences of brain activity at rest are constrained by white matter structure and modulated by cognitive demands.” In: *Communications Biology* 3.1, p. 261. DOI: 10.1038/s42003-020-0961-x.
- Daie, Kayvon, Mark S. Goldman, and Emre R. F. Aksay (Mar. 2015). “Spatial Patterns of Persistent Neural Activity Vary with the Behavioral Context of Short-Term Memory.” In: *Neuron* 85.4, pp. 847–860. DOI: 10.1016/j.neuron.2015.01.006.
- Denisov, D. V. et al. (Feb. 17, 2016). “Universality of slip avalanches in flowing granular matter.” In: *Nature Communications* 7, 10641 EP.
- Eliasmith, Chris et al. (2012). “A Large-Scale Model of the Functioning Brain.” In: *Science* 338.6111, pp. 1202–1205. DOI: 10.1126/science.1225266.
- Faber, Samantha P. et al. (2019). “Computation is concentrated in rich clubs of local cortical networks.” In: *Network Neuroscience* 3.2, pp. 384–404. DOI: 10.1162/netn\_a\_00069.
- Feyerabend, Paul (2010). *Against method*. 4th ed. London ; New York: Verso. 296 pp. ISBN: 978-1-84467-442-8.
- Fiete, Ila R. et al. (Mar. 2010). “Spike-Time-Dependent Plasticity and Heterosynaptic Competition Organize Networks to Produce Long Scale-Free Sequences of Neural Activity.” In: *Neuron* 65.4, pp. 563–576. DOI: 10.1016/j.neuron.2010.02.003.
- Fontenele, Antonio J. et al. (May 2019). “Criticality between Cortical States.” In: *Phys. Rev. Lett.* 122.20, p. 208101. DOI: 10.1103/PhysRevLett.122.208101.
- Fortunato, Santo et al. (Mar. 2, 2018). “Science of science.” In: *Science* 359.6379, eaao0185. DOI: 10.1126/science.aao0185.

- Friedman, Nir et al. (May 2012). “Universal Critical Dynamics in High Resolution Neuronal Avalanche Data.” In: *Phys. Rev. Lett.* 108.20, p. 208102. DOI: 10.1103/PhysRevLett.108.208102.
- Friston, Karl J. (2011). “Functional and Effective Connectivity: A Review.” In: *Brain Connectivity* 1.1, pp. 13–36. DOI: 10.1089/brain.2011.0008.
- Gireesh, Elakkat D. and Dietmar Plenz (2008). “Neuronal avalanches organize as nested theta- and beta/gamma-oscillations during development of cortical layer 2/3.” In: *Proceedings of the National Academy of Sciences* 105.21, pp. 7576–7581. DOI: 10.1073/pnas.0800537105.
- Gu, S et al. (2017). “Optimal trajectories of brain state transitions.” In: *Neuroimage* 148, pp. 305–317.
- Hahn, Gerald et al. (2010). “Neuronal Avalanches in Spontaneous Activity In Vivo.” In: *Journal of Neurophysiology* 104.6, pp. 3312–3322. DOI: 10.1152/jn.00953.2009.
- Haldeman, Clayton and John M. Beggs (Feb. 2005). “Critical Branching Captures Activity in Living Neural Networks and Maximizes the Number of Metastable States.” In: *Phys. Rev. Lett.* 94.5, p. 058101. DOI: 10.1103/PhysRevLett.94.058101.
- Helmer, Markus et al. (Mar. 21, 2017). “Gender bias in scholarly peer review.” In: *eLife* 6, e21718. DOI: 10.7554/eLife.21718.
- Hermundstad, Ann M. et al. (2013). “Structural foundations of resting-state and task-based functional connectivity in the human brain.” In: *Proceedings of the National Academy of Sciences* 110.15, pp. 6169–6174. DOI: 10.1073/pnas.1219562110.
- Honey, Christopher J. et al. (2007). “Network structure of cerebral cortex shapes functional connectivity on multiple time scales.” In: *Proceedings of the National Academy of Sciences* 104.24, pp. 10240–10245. DOI: 10.1073/pnas.0701519104.
- Hopfield, J J (1982). “Neural networks and physical systems with emergent collective computational abilities.” In: *Proceedings of the National Academy of Sciences* 79.8, pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554.



- Ito, Shinya et al. (2011). “Extending Transfer Entropy Improves Identification of Effective Connectivity in a Spiking Cortical Network Model.” In: *PLoS ONE* 6.11. Ed. by Michal Zochowski, e27431. DOI: 10.1371/journal.pone.0027431.
- Jeganathan, J et al. (2018). “Fronto-limbic dysconnectivity leads to impaired brain network controllability in young people with bipolar disorder and those at high genetic risk.” In: *Neuroimage Clin* 19, pp. 71–81.
- Ju, Harang and Danielle S. Bassett (June 2020). “Dynamic representations in networked neural systems.” In: *Nature Neuroscience*. DOI: 10.1038/s41593-020-0653-3.
- Ju, Harang, Jason Z Kim, et al. (2020). “Network structure of cascading neural systems predicts stimulus propagation and recovery.” In: *Journal of Neural Engineering* 17.5, p. 056045. DOI: 10.1088/1741-2552/abbff1.
- Kailath, Thomas (1980). *Linear systems*. Prentice-Hall information and system science series. Englewood Cliffs, N.J: Prentice-Hall. ISBN: 978-0-13-536961-6.
- Kinouchi, Osame and Mauro Copelli (Apr. 23, 2006). “Optimal dynamical range of excitable networks at criticality.” In: *Nature Physics* 2, 348 EP.
- Ko, Ho et al. (Apr. 10, 2011). “Functional specificity of local synaptic connections in neocortical networks.” In: *Nature* 473, 87 EP.
- Kuhn, Thomas S. and Ian Hacking (2012). *The structure of scientific revolutions*. Chicago ; London: The University of Chicago Press. 217 pp.
- Kwak, Haewoon et al. (2010). “What is Twitter, a social network or a news media?” In: *Proceedings of the 19th international conference on World wide web - WWW '10*. the 19th international conference. Raleigh, North Carolina, USA: ACM Press, p. 591. DOI: 10.1145/1772690.1772751.
- Lakatos, Imre (1968). “Criticism and the Methodology of Scientific Research Programmes.” In: *Proceedings of the Aristotelian Society* 69, pp. 149–186.
- Larremore, Daniel B et al. (June 2011). “Effects of network topology, transmission delays, and refractoriness on the response of coupled excitable systems to a stochastic stimulus.” In: *Chaos (Woodbury, N.Y.)* 21.2, pp. 025117–025117. DOI: 10.1063/1.3600760.

- Lefort, Sandrine et al. (2009). “The Excitatory Neuronal Network of the C2 Barrel Column in Mouse Primary Somatosensory Cortex.” In: *Neuron* 61.2, pp. 301–316. DOI: 10.1016/j.neuron.2008.12.020.
- Llinás, Rodolfo R. (Jan. 2003). “The contribution of Santiago Ramon y Cajal to functional neuroscience.” In: *Nature Reviews Neuroscience* 4.1, pp. 77–80. DOI: 10.1038/nrn1011.
- Lydon-Staley, David M. et al. (Mar. 2021). “Hunters, busybodies and the knowledge network building associated with deprivation curiosity.” In: *Nature Human Behaviour* 5.3, pp. 327–336. DOI: 10.1038/s41562-020-00985-7.
- Lynn, Christopher W. and Danielle S. Bassett (May 2019). “The physics of brain network structure, function and control.” In: *Nature Reviews Physics* 1.5, pp. 318–332. DOI: 10.1038/s42254-019-0040-8.
- Markram, H (1997). “A network of tufted layer 5 pyramidal neurons.” In: *Cerebral Cortex* 7.6, pp. 523–533. DOI: 10.1093/cercor/7.6.523.
- Medaglia, J D et al. (2018). “Network Controllability in the Inferior Frontal Gyrus Relates to Controlled Language Variability and Susceptibility to TMS.” In: *J Neurosci* 38.28, pp. 6399–6410.
- Merton, Robert K. (1974). *The sociology of science: theoretical and empirical investigations*. 4. Dr. Chicago: Univ. of Chicago Pr. 605 pp. ISBN: 978-0-226-52092-6.
- Muldoon, S F et al. (2016). “Stimulation-Based Control of Dynamic Brain Networks.” In: *PLoS Comput Biol* 12.9, e1005076.
- Murphy, Kieran A., Karin A. Dahmen, and Heinrich M. Jaeger (Jan. 2019). “Transforming Mesoscale Granular Plasticity Through Particle Shape.” In: *Phys. Rev. X* 9.1, p. 011014. DOI: 10.1103/PhysRevX.9.011014.
- Nagaraj, Abhishek, Esther Shears, and Mathijs de Vaan (Sept. 22, 2020). “Improving data access democratizes and diversifies science.” In: *Proceedings of the National Academy of Sciences* 117.38, pp. 23490–23498. DOI: 10.1073/pnas.2001682117.

- Newman, M. E. J. (Jan. 16, 2001). “The structure of scientific collaboration networks.” In: *Proceedings of the National Academy of Sciences* 98.2, pp. 404–409. DOI: 10.1073/pnas.98.2.404.
- Newton, Isaac (1675). *Isaac Newton letter to Robert Hooke*.
- Nigam, Sunny et al. (2016). “Rich-Club Organization in Effective Connectivity among Cortical Neurons.” In: *Journal of Neuroscience* 36.3, pp. 670–684. DOI: 10.1523/JNEUROSCI.2177-15.2016.
- Nozari, Erfan et al. (Aug. 11, 2021). *Is the brain macroscopically linear? A system identification of resting state dynamics*. arXiv: 2012.12351[cs,eess,math,q-bio].
- Pasqualetti, F, S Zampieri, and F Bullo (2014). “Controllability Metrics, Limitations and Algorithms for Complex Networks.” In: *IEEE Transactions on Control of Network Systems* 1.1, pp. 40–52.
- Perin, Rodrigo, Thomas K. Berger, and Henry Markram (2011). “A synaptic organizing principle for cortical neuronal groups.” In: *Proceedings of the National Academy of Sciences* 108.13, pp. 5419–5424. DOI: 10.1073/pnas.1016051108.
- Petermann, Thomas et al. (2009). “Spontaneous cortical activity in awake monkeys composed of neuronal avalanches.” In: *Proceedings of the National Academy of Sciences* 106.37, pp. 15921–15926. DOI: 10.1073/pnas.0904089106.
- Ponce-Alvarez, Adrián et al. (Nov. 2018). “Whole-Brain Neuronal Activity Displays Crackling Noise Dynamics.” In: *Neuron*. DOI: 10.1016/j.neuron.2018.10.045.
- Popper, Karl R. (1968). *Conjectures and refutations: the growth of scientific knowledge*. Harper torchbooks 1376. New York: Harper. 417 pp. ISBN: 978-0-06-131376-9.
- Robinson-Garcia, Nicolas et al. (Oct. 28, 2020). “Task specialization across research careers.” In: *eLife* 9, e60586. DOI: 10.7554/eLife.60586.
- Rodriguez, Paul and William B Levy (2001). “A model of hippocampal activity in trace conditioning: Where’s the trace?” In: *Behavioral Neuroscience* 115.6, pp. 1224–1238. DOI: 10.1037/0735-7044.115.6.1224.

- Shew, Woodrow L., Wesley P. Clawson, et al. (2015). “Adaptation to sensory input tunes visual cortex to criticality.” In: *Nature Physics* 11.8, pp. 659–663. DOI: 10.1038/nphys3370.
- Shew, Woodrow L., Hongdian Yang, Thomas Petermann, et al. (2009). “Neuronal Avalanches Imply Maximum Dynamic Range in Cortical Networks at Criticality.” In: *Journal of Neuroscience* 29.49, pp. 15595–15600. DOI: 10.1523/JNEUROSCI.3864-09.2009.
- Shew, Woodrow L., Hongdian Yang, Shan Yu, et al. (2011). “Information Capacity and Transmission Are Maximized in Balanced Cortical Networks with Neuronal Avalanches.” In: *Journal of Neuroscience* 31.1, pp. 55–63. DOI: 10.1523/JNEUROSCI.4637-10.2011.
- Shimono, Masanori and John M. Beggs (Oct. 2014). “Functional Clusters, Hubs, and Communities in the Cortical Microconnectome.” In: *Cerebral Cortex* 25.10, pp. 3743–3757. DOI: 10.1093/cercor/bhu252.
- Shriki, Oren et al. (2013). “Neuronal Avalanches in the Resting MEG of the Human Brain.” In: *Journal of Neuroscience* 33.16, pp. 7079–7090. DOI: 10.1523/JNEUROSCI.4286-12.2013.
- Siew, Cynthia S. Q. et al. (June 17, 2019). “Cognitive Network Science: A Review of Research on Cognition through the Lens of Network Representations, Processes, and Dynamics.” In: *Complexity* 2019, pp. 1–24. DOI: 10.1155/2019/2108423.
- Sinatra, R. et al. (Nov. 4, 2016). “Quantifying the evolution of individual scientific impact.” In: *Science* 354.6312, aaf5239–aaf5239. DOI: 10.1126/science.aaf5239.
- Song, Sen et al. (Mar. 2005). “Highly Nonrandom Features of Synaptic Connectivity in Local Cortical Circuits.” In: *PLOS Biology* 3.3. DOI: 10.1371/journal.pbio.0030068.
- Stiso, Jennifer et al. (2019). “White Matter Network Architecture Guides Direct Electrical Stimulation through Optimal State Transitions.” In: *Cell Reports* 28.10, 2554–2566.e7. DOI: 10.1016/j.celrep.2019.08.008.
- Summerfield, Christopher et al. (2006). “Predictive Codes for Forthcoming Perception in the Frontal Cortex.” In: *Science* 314.5803, pp. 1311–1314. DOI: 10.1126/science.1132028.
- Tang, E and D S Bassett (2018). “Control of dynamics in brain networks.” In: *Rev. Mod. Phys.* 90, p. 031003.

- Tavares, Rita Morais et al. (2015). “A Map for Social Navigation in the Human Brain.” In: *Neuron* 87.1, pp. 231–243. DOI: 10.1016/j.neuron.2015.06.011.
- Timme, Nicholas M. et al. (May 2016). “High-Degree Neurons Feed Cortical Computations.” In: *PLOS Computational Biology* 12.5, pp. 1–31. DOI: 10.1371/journal.pcbi.1004858.
- Touboul, Jonathan and Alain Destexhe (Jan. 2017). “Power-law statistics and universal scaling in the absence of criticality.” In: *Phys. Rev. E* 95.1, p. 012413. DOI: 10.1103/PhysRevE.95.012413.
- Wang, Dashun, Chaoming Song, and Albert-László Barabási (Oct. 4, 2013). “Quantifying Long-Term Scientific Impact.” In: *Science* 342.6154, pp. 127–132. DOI: 10.1126/science.1237825.
- Wang, Xiao-Jing (Sept. 2002). “Probabilistic Decision Making by Slow Reverberation in Cortical Circuits.” In: *Neuron* 36.5, pp. 955–968. DOI: 10.1016/S0896-6273(02)01092-9.
- Wang, Yun, Henry Markram, et al. (Mar. 19, 2006). “Heterogeneity in the pyramidal network of the medial prefrontal cortex.” In: *Nature Neuroscience* 9, 534 EP.
- Watts, Duncan J. and Steven H. Strogatz (June 4, 1998). “Collective dynamics of ‘small-world’ networks.” In: *Nature* 393, 440 EP.
- Wibral, Michael, Conor Finn, et al. (2017). “Quantifying Information Modification in Developing Neural Networks via Partial Information Decomposition.” In: *Entropy* 19.9, p. 494. DOI: 10.3390/e19090494.
- Wibral, Michael, Viola Priesemann, et al. (Mar. 2017). “Partial information decomposition as a unified approach to the specification of neural goal functions.” In: *Brain and Cognition* 112, pp. 25–38. DOI: 10.1016/j.bandc.2015.09.004.
- Wilting, J and V Priesemann (2019). “25 years of criticality in neuroscience — established results, open controversies, novel concepts.” In: *Current Opinion in Neurobiology* 58, pp. 105–111. DOI: 10.1016/j.conb.2019.08.002.
- Wilting, Jens and Viola Priesemann (2018). “Inferring collective dynamical states from widely unobserved systems.” In: *Nature Communications* 9.1, p. 2325. DOI: 10.1038/s41467-018-04725-4.

- Wu, Lingfei, Dashun Wang, and James A. Evans (Feb. 2019). “Large teams develop and small teams disrupt science and technology.” In: *Nature* 566.7744, pp. 378–382. DOI: 10.1038/s41586-019-0941-9.
- Zeng, An et al. (Nov. 2017). “The science of science: From the perspective of complex systems.” In: *Physics Reports* 714-715, pp. 1–73. DOI: 10.1016/j.physrep.2017.10.001.

## CHAPTER 2

### DYNAMIC REPRESENTATIONS IN NEURAL NETWORKS

*This chapter contains work from Ju, H. and Bassett, D.S. (2020). “Dynamic representations in networked neural systems.” Nature Neuroscience 23, 908–917.*

#### 2.1. Abstract

A group of neurons can generate patterns of activity that represent information about stimuli; subsequently, the group can transform and transmit activity patterns across synapses to spatially distributed areas. Recent studies in neuroscience have begun to independently address the two components of information processing: the representation of stimuli in neural activity and the transmission of information in networks that model neural interactions. Yet only recently are studies seeking to link these two types of approaches. Here we briefly review the two separate bodies of literature; we then review the recent strides made to address this gap. We continue with a discussion of how patterns of activity evolve from one representation to another, forming dynamic representations that unfold on the underlying network. Our goal is to offer a holistic framework for understanding and describing neural information representation and transmission while revealing exciting frontiers for future research.

#### 2.2. Introduction

Organisms live in and interact with an ever-changing environment. From the microscopic nematode searching for single molecules of food to a cat readying to pounce on a mouse, animals must gather sensory cues to identify environmental variables that could either aid or end their survival. Is this berry edible or poisonous? Is that the sound of a hawk waiting to stoop or just the sound of wind through grass? Even the simplest of these questions necessitates an internal representation of one’s environment and how that environment may

change over time. Since the mid-20th century, experiments have used diverse sensory stimuli and tasks to elucidate the scope of neural representations, from mapping social hierarchies to predicting future events that impact survival (Summerfield et al., 2006; Tavares et al., 2015).

More recently, developments in empirical methods and theory have allowed the quantification of not only static neural representations, but also dynamic representations that evolve appreciably in time (Gallego et al., 2018). In these dynamic representations, neural activity follows spatiotemporal patterns that are associated with complex, dynamic abstractions, such as remembering sequences of visual patterns or motor control (Gallego et al., 2018; Shine et al., 2019; Chaudhuri et al., 2019). With this new knowledge come new questions about the dynamic nature of neural representations: what kinds of dynamic abstractions can neurons represent? And how and why do representations change over time to support behavior?

In this Review, we posit that a fundamental understanding of neural representations may lie in understanding the networks of interactions between neural units (Bassett and Sporns, 2017). Neural representations are thought to arise from patterns of neuronal firing (Saxena et al., 2019); importantly, neurons do not fire in isolation. Rather, they are intricately connected within a complex network of synapses on which activity propagates from one neuron to another. By abstracting complex interactions, we can use network models, dynamical systems theory and other approaches to understand the behaviors that emerge from neural systems.

To begin, we briefly review recent work from the two fields of neural representations and network models before describing efforts to bridge them. Further innovation, however, will require new theoretical and methodological developments. Thus, we outline a general theoretical framework for reasoning about the dynamics of representations in networked neural systems. To build this framework, we gather recent theories and evidence that support mechanisms by which representations evolve over time, through intraregional dynamics and



the interactions between intra- and inter-regional dynamics. Our hope is that by bridging the two fields, we will facilitate the quantification and explanation of dynamic representations, which in turn will open doors to a deeper understanding of the neural computations that underlie cognition in complex, dynamic environments.

### **2.3. Neural representations**

Neural activity can represent a variety of physical and abstract variables in the environment. For example, neurons in the hippocampus and entorhinal cortex can selectively activate in response to spatial cues, such as the egocentric location of an animal or of others (Danjo et al., 2018). Neurons in the same area are also selectively responsive to conceptual knowledge, such as the shapes of objects (Constantinescu et al., 2016) and the social environment (Tavares et al., 2015).

In studying neural representations, an important recent step has been to measure how a population of neurons or voxels (i.e., volumes of brain tissue) can represent variables by activating in a specific spatial pattern in response to a particular stimulus pattern (Saxena et al., 2019). The encoding of representations in neural populations offers a computational advantage over encoding in individual neurons, especially in complex cognitive tasks (Rigotti et al., 2013; Parthasarathy et al., 2017). By observing neural populations, studies have demonstrated that neurons can represent abstract phenomena, such as visual objects (Rigotti et al., 2013; Stringer, Pachitariu, Steinmetz, Carandini, et al., 2019), events (Schapiro et al., 2013), tasks (Yang et al., 2019), social cues (Levy et al., 2019), and language (Arana et al., 2020) (Figure 2.1A). Even some trial-by-trial fluctuations in neural activity that were once considered statistical noise are now known to be shaped by an animal’s various physical movements (Musall et al., 2019; Stringer, Pachitariu, Steinmetz, Reddy, et al., 2019). Whereas representations of simple, physical variables like spatial location can support survival by guiding immediate responses to the environment, more abstract representations can be important in building a richer model of the world, which can support survival over

longer timescales through prediction and planning (Mobbs et al., 2020).

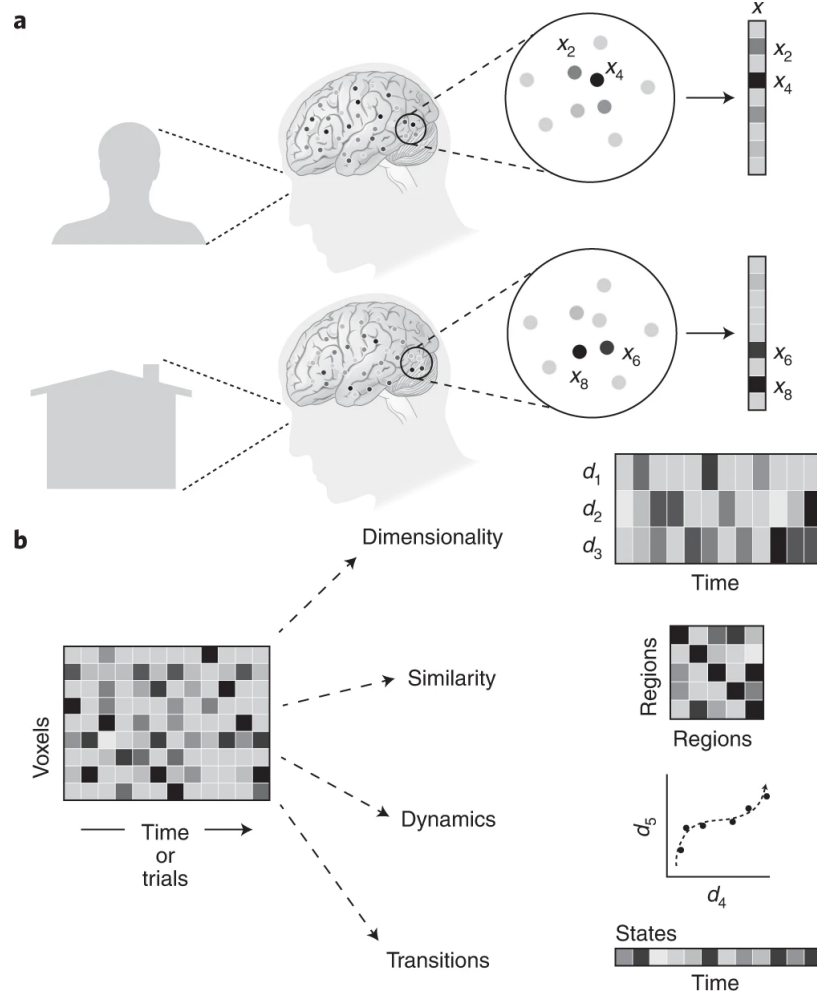


Figure 2.1: **Neural representations and tools to analyze them.** **a** An illustration of an experimental setup for studying neural representations. A participant views faces (top) and objects (bottom) while an experimenter records the subject’s neural activity. The shade of each point indicates the activity level of voxels, and voxels can be grouped into a population, whose activity can then be represented as a vector,  $x$ . **b** Methods to analyze neural activity, as a time-series or by trials: dimensionality reduction methods identify dimensions (typically using two or three for visualization and interpretation;  $d_1$ ,  $d_2$ , and  $d_3$ ) that explain a large portion of the observed variance in neural activity; similarity analysis methods compare activity patterns of voxels between brain regions or experimental conditions; dynamical methods analyze the change in activity patterns, for example, how the activity patterns of one population,  $d_4$ , change as a function of the activity patterns of another population,  $d_5$ ; and other methods quantify how one type of activity pattern can transition to another type, with each type being referred to as a ‘state’.

To study representations in neural populations, one can operate on the single variable of population-averaged activity or one can operate on a vector of neuronal activity within a population (Saxena et al., 2019) (Figure 2.1B). Studies of population activity often employ principal component analysis (PCA) (Gallego et al., 2018; Shine et al., 2019; Parthasarathy et al., 2017; Stringer, Pachitariu, Steinmetz, Carandini, et al., 2019; Levy et al., 2019) or linear models (Musall et al., 2019; Tang et al., 2019) to quantify the inherent dimensionality of the dynamics. Interestingly, such methods and related techniques (McIntosh et al., 2013) show that the space of population activity can be either high-dimensional or low-dimensional. At high dimensionality, population activity encodes information more efficiently, as in encoding visual stimuli (Parthasarathy et al., 2017; Stringer, Pachitariu, Steinmetz, Carandini, et al., 2019), whereas at low dimensionality, activity encodes more robustly, as in complex cognitive or motor tasks (Tavares et al., 2015; Gallego et al., 2018; Tang et al., 2019). More modern multivariate methods, such as representational similarity analysis (RSA) (Kriegeskorte, 2008) and multivoxel pattern analysis (MVPA) (Mahmoudi et al., 2012), abstract representations away from precise activity patterns in favor of focusing on the similarities between patterns across experimental conditions characterized by distinct stimuli or tasks. Collectively, these multivariate methods capture neural representations of population activity in individual brain regions.

Yet the question remains: how do neurons or larger neural units form, change and transmit representations? To answer this question, a key observation may be the temporal component of neural representations: neural activity evolves over time to represent dynamic variables (Figure 2.1B). In the theory of hippocampal sequence learning, temporal patterns of activity in hippocampal cell assemblies encode sequences of locations (Dragoi et al., 2006; Epstein et al., 2017) and episodic memory through oscillatory activity (Lisman et al., 1995). More complex trajectories of system-wide neural activity serve higher cognitive functions, for instance, in motor and cognitive tasks (Tavares et al., 2015; Gallego et al., 2018; Taghia et al., 2018), sometimes measured as transitions between discrete moments of activity, typically called states (Cornblath et al., 2020) (Figure 2.1B). The dynamic nature of neural

representations prompts further discussion of how representations can evolve on a network of synapses or white matter tracts that connect neurons and brain areas.

## 2.4. Network models

To understand how neural representations evolve over time, we propose that the dynamic evolution of representations emerges from the interactions between neural units. Thus, we briefly explore network models in neuroscience before discussing in the next section how they may give rise to representations. Neurons (and neuron populations) are intricately connected in a complex web of interactions, and network models abstract neural units and their connections as a network of nodes and edges (Bassett, Zurn, et al., 2018). The pattern, or topology, of these connections—within and across neurons, populations of neurons and ultimately brain regions—constrains activity in complex, dynamical neural systems.

To build a network model of the brain, one can quantify either the structural connections or the dynamic interactions between neurons or brain regions (Figure 2.2). In humans, structural connections are often estimated from the diffusion of water along white matter tracts that connect distant brain areas (Johansen-Berg, 2013). In contrast, dynamic interactions are reflected in the effective connectivity of a neural system, which describes the putatively causal interactions between neural units (Friston, 2011). Approaches to estimate effective connectivity include linear autoregressive models (Friston, 2011; Neumaier et al., 2001), information theoretic measures of transfer entropy (Ito, Hansen, et al., 2011) and probabilistic Bayesian models of dynamic causal modeling (Friston et al., 2003). Network models of both structural and effective connections inform how patterns of connections may mediate the dynamic processes that flow on top of these connections (Bassett and Sporns, 2017).

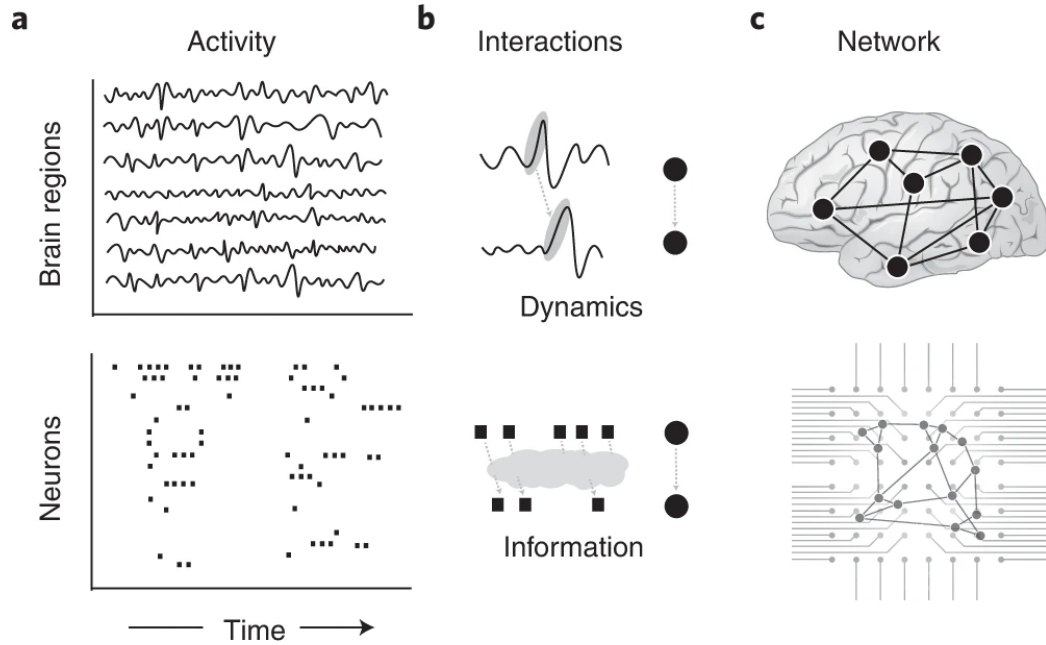


Figure 2.2: **Network models abstract neural systems.** **a** The activity of multiple brain regions (top) or neurons (bottom), as time-series. **b** Model-based (top) or information theoretic (bottom) methods capture the time-lagged interactions between pairs of nodes. Such models can be informed by structural links between node pairs, such as synapses between neurons or white matter tracts between large-scale brain areas. **c** A network model is constructed from the pairwise functional or effective interactions, whose estimation is depicted in B, or from pairwise structural connections (not shown).

By applying recent methods from network neuroscience (Bassett, Zurn, et al., 2018), one can quantitatively characterize global, mesoscale and local patterns of connectivity in brain networks (Bassett and Sporns, 2017; Bassett, Zurn, et al., 2018). For example, many empirical networks, including brain networks, display global architectures that lie in between those of random and ordered networks, in a manner that is well-described by the Watts–Strogatz small-world model (Bassett and Bullmore, 2006). Mesoscale architecture can be reflected in modular and core–periphery structures (Sporns et al., 2016; Rombach et al., 2014), whereas local architecture can be reflected in hubs, which can join together to form ‘rich clubs’ (Heuvel and Sporns, 2011; Bullmore et al., 2013). Collectively, network measures distill the complex patterns of connections down to simple organizing principles across topological and spatial scales.

Network models have yet to give much insight into neural representations and, ultimately, cognition—despite the fact that such models enhance our understanding of the characteristics of information transmission in the brain (Summerfield et al., 2006; Avena-Koenigsberger et al., 2018). Initial efforts suggest that short paths characteristic of small-world networks (Bassett and Bullmore, 2006) together facilitate the spread of signals throughout a network (Avena-Koenigsberger et al., 2018; Mišić et al., 2015). Similarly, rich clubs of local cortical neurons propagate and process information (Faber et al., 2019). Other topological features, such as the topological similarity between two regions, can predict functional correlations in their activity (Bettinardi et al., 2017). These initial efforts underscore the potential for network models to contribute much more toward our understanding of how neural representations evolve and support cognition.

## 2.5. Integrating neural representations and network models

While neural representations relate environmental or behavioral variables to neural activity, network models estimate and predict changes in neural activity, and recent studies have begun to integrate neural representations and network models. In particular, new methods estimate inter-regional dynamical interactions using statistical relationships between activity patterns measured from functional MRI (Kriegeskorte, 2008; Anzellotti et al., 2017) (Figure 2.3A). As representations are transmitted from one brain region to another, one can quantify how they are transformed using a linear model (Anzellotti et al., 2017) or similarity analysis (Kriegeskorte, 2008) (Figure 2.3B). Other multivariate methods, such as multivariate pattern dependence (MVPD; versus the MVPA mentioned earlier), can tease apart features of representations in a brain region, such as the low- versus high-level properties of faces in the fusiform, which are differentially transmitted to disparate brain regions (Anzellotti et al., 2017). These and other mathematical methods, such as sheaves from algebraic topology (Curry, 2014), can be applied to neural data to inform our understanding of how activity patterns change as they are transmitted across brain regions.

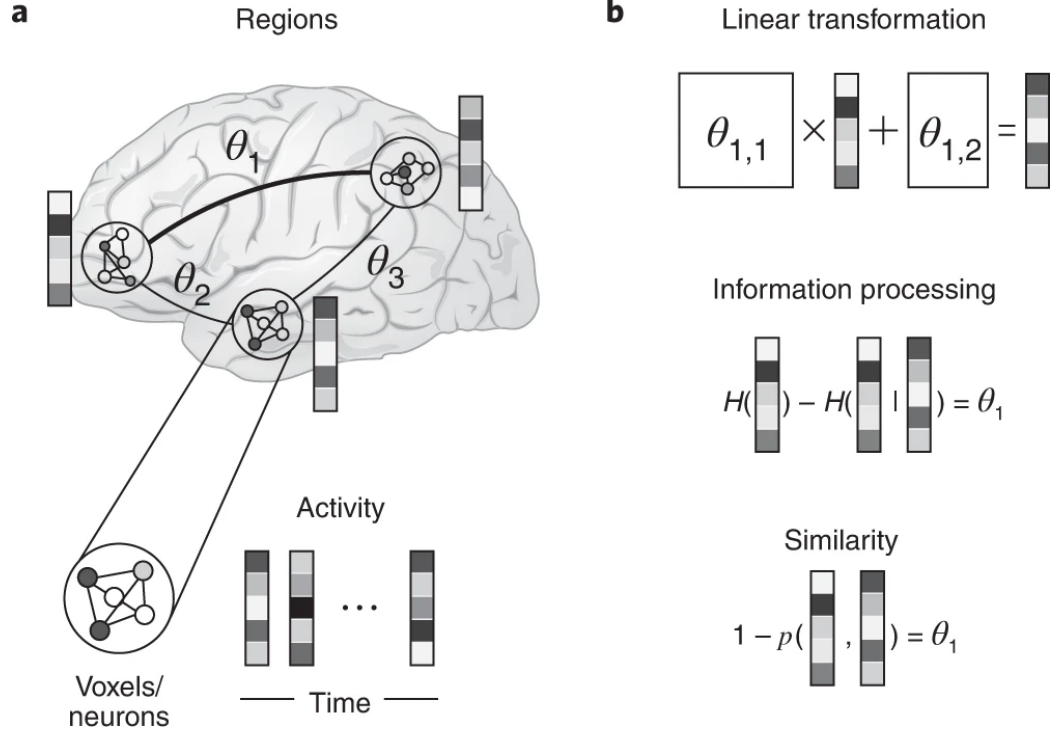


Figure 2.3: **Integrating network models and neural representations.** **a** a, Network models abstract the interactions between brain regions with connections  $\Theta = \Theta_1, \dots, \Theta_N$ , where each connection  $\Theta_i$  may be composed of  $M$  parameters  $\Theta_{i,1}, \dots, \Theta_{i,M}$ . **b** Recent efforts examine the transmission of representations across brain regions as transformations of activity, transmission of information ( $H$  indicates entropy) or similarity in activity ( $\rho$  indicates correlation).

In a similar line of inquiry, others have sought to quantify the information carried across brain regions (Ito, Hearne, et al., 2020). Multivariate methods, such as information connectivity or information transfer mapping, measure this information as the synchrony in the discriminability of multivariate patterns (Coutanche et al., 2013). The information carried in multivariate patterns can also be estimated using classical information-theoretic measures, such as mutual information, of spatial and temporal multivariate patterns (Shannon et al., 1998; Ju et al., 2020). At the cellular level, empirical studies have measured non-linear, information-theoretic dependencies between neurons (Figure 2.3B) (Ito, Hansen, et al., 2011). Using partial information decomposition, an even more recent approach, one can measure the shared, unique and synergistic transmission of information across neural

networks (Faber et al., 2019; Wibral et al., 2017).

What is the mechanism by which such transformation and transmission of information occurs? Anatomical pathways are a key candidate. Structural connections are robustly linked with correlations in activity (Bansal et al., 2018; Hermundstad et al., 2013) and dynamics (Cornblath et al., 2020). By approximating linear dynamics on those connections, one can provide closed-form, analytic solutions to dynamical properties, such as the minimum energy to control the activity of one region from another (Kim et al., 2018), with numerical approximations also informing potential clinical applications (Stiso et al., 2019). However, the linear approximation may be less appropriate higher in the cortical hierarchy, where the link between structure and activity correlations diverges (Vázquez-Rodríguez et al., 2019). These and related studies suggest the need to understand both simple (linear) and more complex mechanisms by which inter-regional dynamics arise from interregional structural connections.

Efforts to meet that need will benefit from the multivariate methods discussed in the previous section, such as RSA and MVPA, which reveal that the multivariate activity patterns within a brain area can flexibly represent environmental and task-relevant variables. Moreover, they would benefit from the new methods discussed in this section, which show ways to estimate multivariate dynamics across brain areas. Indeed, recent work highlights the importance both of the dynamics of representations within a brain area and the constraints imposed on those dynamics by the underlying network. Theoretical studies have already examined how the circuitry within a neural population can support a variety of computations (Wolf et al., 2014; Weber et al., 2019), such as Bayesian computation (Sohn et al., 2019). Thus, understanding how neural networks form, change and transmit representations in the brain at various scales seems fundamental to understanding the computations that underlie cognition.



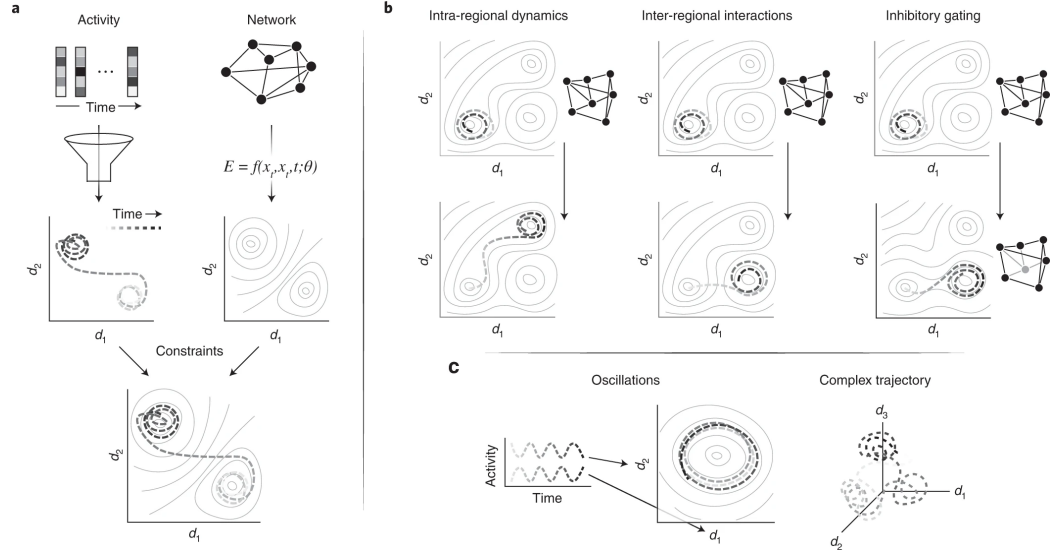
## 2.6. Dynamic representations in networked neural systems

Looking forward, we posit two topics of research that may facilitate the further integration of neural representations and network models: (1) intraregional dynamics and (2) the interactions between intra- and inter-regional dynamics (Figure 2.4 and **Box 1**). These two topics can inform how representations are formed, changed and transmitted along network connections. We review disparate bodies of literature that support the integration of representations and networks and that further hint at the insights potentially gleaned therefrom. Then, in the next sections, we propose a theoretical framework for dynamic neural representations in networked neural systems and review methods that have already (and not yet) been applied to study real neural systems.

### 2.6.1. Intraregional dynamics

For the purposes of our discussion here, we generalize the term ‘region’ to refer to a collection of neighboring neural units, from neurons to voxels. Moreover, we draw a distinction between the macroscopic scale of whole-brain imaging and the microscopic scale of neural recordings, such as those using local field potentials (LFPs). At the macroscopic scale, the multivariate methods discussed previously, such as RSA and MVPD, operate on brain areas. At the microscopic scale, a region may refer to, for example, a few hundred neurons (Heuvel, Stam, et al., 2008). Considering the modular yet interconnected nature of brain networks (Sporns et al., 2016; Bertolero et al., 2015), we wish to distinguish the interactions among subsystems, which we refer to as brain regions, from the interactions within subsystems. Thus, we will first discuss how a region evolves through intraregional dynamics, how it evolves autonomously without inter-regional interactions, and then how it interacts with other regions through inter-regional connections.

It is difficult to empirically observe autonomous, intraregional brain dynamics because each region is continually interacting with other regions. The stronger the interactions between brain regions, the further the interactions drive down the dimensionality of the observed



**Figure 2.4: Dynamic representations in networked neural systems.** **a** Neural activity (left) dynamically evolves along a trajectory (dashed) in a system space with dimensions  $d_1$  and  $d_2$ . Neural network models (right) constrain dynamics by parameterizing an energy function, thus constructing an energy landscape (illustrated by contour lines, where circular lines indicate energy wells; for illustration purposes, we assume a smooth energy landscape). Network models can inform the description, prediction and control of dynamic neural representations. **b** Dynamics of neural representations in networks (arrows indicate time). In autonomous, intraregional dynamics (left), network states travel along valleys. Through inter-regional interactions (middle), network states can travel over hills. Reconfiguration of network dynamics (right), through for example inhibition of a node (greyed), creates new hills and valleys in the landscape. **c** Examples of observations of dynamics: oscillations (left) and more complex trajectories (right). See **Box 2** for details.

dynamics, to a point where one or two dimensions can explain most of the variance in neural activity. In whole-brain imaging during motor tasks, high-dimensional cortical activity converges onto a low-dimensional manifold in mice (Chaudhuri et al., 2019) and monkeys (Gallego et al., 2018). Despite such convergence, cortical activity can still be spatially decomposed into linear kernels that represent uninstructed movements (Musall et al., 2019) or into orthogonal principal components that are then integrated in performing cognitive and motor tasks (Shine et al., 2019).

We briefly demonstrate the mathematical dependence of dimensionality of activity on the decomposability of the network structure. If, for simplicity, we approximate neural interactions as linear functions, then we can use spectral decomposition methods to find modular

subsystems, each with an eigenvector with a large associated eigenvalue (Kailath, 1980). If, for example, the network has two modules and dynamics are approximately linear, then around two principal components (i.e., eigenvectors of the network’s covariance matrix) will explain most of the variance in activity. However, in the contrasting case in which network structure is complex or dynamics highly nonlinear, then low-dimensional representations are unlikely to accurately account for the system’s function.

Dimensionality-reduction approaches that are commonly applied to data from multiple regions can also be used to summarize intraregional dynamics. At the macroscale, for example, one can apply dimensionality analysis to isolate the endogenous, intraregional dynamics of individual brain regions at rest (Hermundstad et al., 2013; Fox et al., 2005) and at differing levels of sensory deprivation (Chang et al., 2016). These data summaries of the observed dynamics can be complemented by network models explaining how those dynamics arise from intraregional architecture. Voxels within brain regions, for example, show non-random correlational structures (Heuvel, Stam, et al., 2008) that could inform network models built to make predictions about intraregional dynamics that are relevant for perception, cognition or behavior (Figure 2.4A).

At the microscopic scale, intraregional dynamics can manifest in spontaneous cascades of spikes in cortical tissues, often observed *in vitro* and *in vivo* (Stringer, Pachitariu, Steinmetz, Reddy, et al., 2019; Beggs, 2004). In such neuronal avalanches, neural populations can operate between regimes of activity decay and amplification (Wilting et al., 2019), and they can be probed further to reveal the relationship between network structure and properties of intraregional dynamics (Ju et al., 2020). Notably, spontaneous cascades robustly follow specific patterns of activity (Beggs, 2004), in which the neural interactions can be mapped as a network (Ito, Hansen, et al., 2011) to make predictions about the system’s dynamics (Figure 2.4B) (Ju et al., 2020). These cascades have been observed even in *ex vivo* experiments performed on turtle brains (Shew et al., 2015), in which one can measure intraregional dynamics before and after stimulating a separate but interacting brain region

to observe the relationship between intra- and inter-regional dynamics.

Perhaps periodic cycles are one of the simplest examples of dynamics that are intrinsic to the neural interactions across network connections. For example, central pattern generators display attractor-like dynamics that are stable around a periodic cycle (Kailath, 1980). The gastric mill circuit in crustaceans is particularly well characterized; the circuit consists of four ganglia whose networked interactions produce very specific patterns of activity that allow crustaceans to feed (Nusbaum et al., 2002). More generally, oscillations in neural systems also undergo periodic attracting behavior (Figure 2.4C) (Buzsáki et al., 2004). These cortical oscillations are observed in individual brain regions (Bartos, Vida, et al., 2007), and their synchrony across brain regions has been linked to cognitive processes, such as attention and memory (Joo et al., 2018).

### **2.6.2. Theoretical framework**

Thus far, we have reviewed the dynamics of neural representations within and between brain regions that may give rise to cognition. But how can network models help us better understand those dynamics? We propose that they help us understand constraints on system dynamics (Figure 2.4A). To explore this idea, we now outline a general theoretical framework for dynamic neural representations built from recent empirical and theoretical literature. The driving principle behind this framework is that network structure constrains the manner in which a system evolves, or transitions, through a pattern of states (Bassett, Zurn, et al., 2018). These transitions link together different states into a sequence that can represent dynamic variables.

To visualize the mechanisms of state transitions, consider the energy landscape of a neural system (Figure 2.4A) (Gu, Cieslak, et al., 2018). As in an actual topographic map, this energy landscape can have hills that are difficult to reach and valleys toward which it is easy to descend. Mathematically, we can represent the activity pattern of a neural system as the configuration state of all the neural units (for example, neurons, voxels or regions).

At a particular time  $t$ , this state is denoted by the vector  $x_t$  and has an associated energy, which formalizes the activity in a system (see Box 1 for a glossary of terms) (Kailath, 1980). The system also has network parameters  $\Theta$  that mediate the interactions between neural units and thus the difficulty with which a system can go from  $x_t$  in a direction  $\dot{x}_t$ . Thus, the energy landscape can be formulated as  $E = f(x_t, \dot{x}_t, t; \Theta)$ , parameterized by the network  $\Theta$  (Figure 2.4A) (Hopfield et al., 1986). Importantly, this formulation reduces the  $n$ -dimensional state space and  $n^2$ -dimensional network connection space into a one-dimensional energy function, as often used in network control theory (Kim et al., 2018; Gu, Pasqualetti, et al., 2015). The system state traverses this energy landscape, much like a traveler traverses a landscape of hills and valleys or, as is relevant to our discussion, the mind traverses a landscape of cognitive possibilities.

### 2.6.3. Inter-regional interactions

Regions interact with one another to influence their individual dynamics. While intraregional dynamics have been studied in the context of similarity and pattern dependence, we now use our theoretical framework to consider how regions may interact with one another (Figure 2.4B). Input from one brain region to another has been primarily studied in the context of system control (Liu et al., 2011). The idea of system control has roots in the cognitive control hypothesis, which states that higher-level processing regions exert executive control over the states of lower-level regions (Cools et al., 2019; Badre et al., 2019) to, for example, selectively attend to a stimulus (Lavie et al., 2004).

More recently, studies have adapted control theory to networked systems, including brain networks, to predict which brain regions can effectively control brain activity (Kim et al., 2018; Gu, Pasqualetti, et al., 2015; Yan et al., 2017). Dynamical systems theory provides mathematical insight into a system’s dynamical properties, such as the stability of a swinging pendulum, given an approximate model of the system (Kailath, 1980). Through this control framework, Yan et al. (2017) predicted which neurons control what locomotor behaviors

in the connectome of the nematode *Caenorhabditis elegans* using a linear approximation of neural interactions. This framework can also be applied to indirect, neuroimaging measurements, where the multivariate interactions between brain regions (Kriegeskorte, 2008; Anzellotti et al., 2017) can be modeled as one brain region receiving input, or control, from another brain region (Kim et al., 2018). Even sensory input and brain–computer interfaces can potentially be thought of as control mechanisms, in which the whole brain is driven by sensory or artificial stimulation (Hatsopoulos et al., 2009). According to this framework, through control across brain regions, a system can follow state trajectories that are normally not accessible to autonomous intraregional dynamics.

Through inter-regional interactions, one region can not only somewhat control the state of another region along a trajectory, but also change the nature of the trajectory itself (Figure 2.4B) (Katz et al., 1996). Consider a simple example of a network with a neuron that receives inhibitory input from an external neuron. To reach a state in which the neuron is active, the network would normally require little energy. However, when the neuron receives inhibitory inputs, the neuron can no longer spike, and the network now requires an exorbitant amount of energy to reach the same state (Barbas et al., 2007). Such gating can modulate the interactions between intrinsic dynamics and sensory inputs or inter-regional feedback (Stringer, Pachitariu, Steinmetz, Okun, et al., 2016; Eschbach et al., 2020; Bartos, Manor, et al., 1999) and thus induce complex dynamics through reconfiguration of the energy landscape.

In line with this notion of malleability of the energy landscape, recent studies have suggested that the inter-regional correlation structure of neural systems can be dynamically reconfigured. In whole-brain imaging, many studies have observed that the functional networks, which measure correlation across brain regions, change during tasks (Cohen et al., 2008; Bassett, Wymbs, et al., 2011; Kilteni et al., 2020), from motor skill learning to narrative comprehension. Through these dynamic changes in functional networks, information can be flexibly routed across neural units (Avena-Koenigsberger et al., 2018; Palmigiano et al.,

2017; Kirst et al., 2016). On the cellular scale, gating mechanisms can explain how a network can dynamically direct activity or information through various neural circuits, such as those involved in fear learning (Krabbe et al., 2019; Cummings et al., 2020), vocalization (Tschida et al., 2019), and locomotion (Clancy et al., 2019). Indeed, the ‘gate’ in gating theory refers to a gate-like hill in an energy landscape. Many models use inhibitory circuits to dynamically gate the flow of activity (Coleman et al., 1995; Popov et al., 2018) via shunting inhibition, which effectively turns off the postsynaptic neurons (Borg-Graham et al., 1998). These studies could be related to dynamic changes in the energy landscape that are determined by the underlying network structure. In addition, neural gain theories suggest that the locus coeruleus–norepinephrine system (Aston-Jones et al., 2005) may increase average activity in groups of neurons, thereby allowing the flexible control of functional connectivity networks (Haider et al., 2009) and the modulation of attention and learning (Eldar et al., 2013). Through gating or gain control, neurons external to a population can drastically modulate the energy landscape, turning energy hills to energy valleys or vice versa.

#### **2.6.4. Applications to real neural systems**

Despite the recent progress, studies have yet to more concretely link the dynamics within and across neural populations to cognition. By reviewing the theory and evidence surrounding this gap in knowledge, we wish to outline current methods and to motivate new research on how a network constrains the dynamics for neural activity within and across brain regions. While the mathematical generality of the theoretical framework lends itself to a holistic description of brain dynamics (**Box 2**), it also requires the introduction of assumptions and approximations to begin to be applied to real neural systems. Here we discuss such methods that have already been applied to real and theoretical neural systems, as well as methods from other fields that have yet to be applied to neural data. Using these methods, studies have and can give insight into the dynamic, networked underpinnings of cognition.

To model the dynamics of neural interactions across a large population of neurons, one can

use approximations of neural dynamics that are computationally efficient or even analytically solvable. The simplest approximation is that of a linear, time-invariant system. One can efficiently approximate the dynamics of such systems using vector autoregression, both on functional MRI data (Friston, 2011) and spiking data (Ju et al., 2020), or through subspace system identification (Becker et al., 2018). After building a networked, dynamical model, one can use the model to make predictions about system behavior, including the activity signals themselves (Becker et al., 2018). In a framework of network control theory, a network of structural connections between brain regions can predict the stability of activity patterns in neuroimaging measurements during the  $N$ -back working memory task, and the stability is modulated by D1 and D2 dopamine receptors (Nozari et al., 2021). Such linear models can also predict a neuron’s role in controlling motor actions within the *C. elegans* connectome (Yan et al., 2017) and the spread of electrocorticography (ECoG) stimulation through white matter tracts (Stiso et al., 2019). In a bipartite, inter-regional brain network with linear dynamics, analytical solutions exist to the minimum energy required to control one region from another (Kim et al., 2018).

Linear, time-invariant network models can be extended with nonlinearities or stochasticity to better model the complex behavior of neural systems. For example, in linear-threshold models, analytical solutions also exist to predict how neurons may attend selectively to stimuli (Nozari et al., 2021). A dynamical hybrid system is another concept from dynamical systems theory that has yet to be applied to neural data. Interestingly, a hybrid system evolves continuously with discrete, stochastic ‘jumps’ between state trajectories, like a bouncing ball that rises and falls in an arc but exhibits inelasticity in the collision with a surface (Henzinger, 1996). Such systems could prove useful in describing brain regions that perform multiple ‘functions’, made possible in part by inhibitory circuits (Coleman et al., 1995; Popov et al., 2018).

Both linear and nonlinear models of network dynamics can be strengthened by more accurate estimates of the networks themselves. One can, for example, build networks from either



high-resolution structural connections or high-precision effective connections to obtain increasingly accurate predictions of neural activity, as we have discussed previously. To more precisely map structural connections, some have used subsampling of the shortest structural paths between individual cortical voxels (Greene et al., 2019). The precise mapping between cortical regions can help identify the specific inter-regional networks that underlie a range of neurologic symptoms (Fox, 2018). Beyond simple region-to-region connections, sheaves from algebraic topology have yet to be applied to neural data, but can determine the formal mathematical maps between the multivariate activity of nodes and the inter-regional interactions of edges (Curry, 2014).

The more detailed network models accessible to these approaches can be used within large-scale dynamical models of brain activity to further inform the integration of representations and networks by multivariate methods. While multivariate methods like RSA and MVPD can be extended to describe the transmission of representations across brain areas, the exact neural dynamics that underlie that transmission of representations remains poorly understood. What could complement these empirical methods are large-scale models of brain dynamics (Breakspear, 2017), which have been shown to exhibit neural and cognitive behaviors resembling those observed empirically, including propagating waves of brain activity (Roberts et al., 2019) and a copy-and-drawing cognitive task (Eliasmith et al., 2012). We posit that large-scale, dynamical network models of neural representations can be used hand-in-hand with burgeoning empirical methods to reveal the neural mechanisms that give rise to perception and cognition.

### **2.6.5. Dynamic representations in cognition**

Finally, we review important cognitive constructs that may benefit from the framework of dynamic neural representations. Perhaps the most fundamental representation of an organism’s environment is one that describes the physical world in the dimensions of space and time. Early theories of cognition describe how animals form representations of physical

space (Danjo et al., 2018). Along another line of research, early theories describe neural representations of physical shapes (Marr et al., 1978), and recent empirical literature reveals how the brain gathers and processes visual information to represent an object invariant to many features of such representations (DiCarlo et al., 2012).

More recent evidence supports the idea that the brain navigates through the physical world similarly to the way it navigates through more abstract constructs (Epstein et al., 2017). In episodic memory, navigation through physical space and cognitive abstractions occur (Tavares et al., 2015; Constantinescu et al., 2016) through neural mechanisms in the prefrontal cortex and entorhinal cortex–hippocampal subsystems (Ekstrom et al., 2003). Attention can be framed in a similar way. Indeed, in episodic memory, one may be attending to, or imagining (Hassabis et al., 2007), event representations of past experiences (Richmond et al., 2017; Brunec et al., 2018), while attention is conventionally viewed as attending to representations of the immediate environment (Chun et al., 2011). The complementary learning systems theory offers a similar perspective, in which the cortical system learns structured ‘items’ of knowledge while the hippocampal system rapidly learns the relationships, temporal and otherwise, between items (Kumaran et al., 2016).

While our current understanding of physical and abstract representations is as yet incomplete, the framework of dynamic neural representations motivates the study of how various neural subsystems—whether neural populations or brain regions—represent aspects of the physical world and how the subsystems interact to represent the dynamics of the physical world. This line of investigation has been driven in part by research in core human knowledge (Spelke et al., 2007). For example, intuitive physics (Battaglia et al., 2012) studies how infants have an intuitive understanding of basic physics, such as that of falling objects (Téglás et al., 2011). There exist models of intuitive physics, such as those in machine learning, but how the brain performs this complex though intuitive task has yet to be understood (Kumaran et al., 2016). How does the brain form, change and ultimately extinguish the representations of objects? By studying representations of objects with modern multivari-

ate methods and how these representations evolve within a brain region and interact with other brain regions, one can begin to understand the neural dynamics that support intuitive physics and potentially more abstract cognitive tasks.

Lastly, integral to the framework is not only the dynamics within brain regions but also the interactions of dynamics between regions, which can be modeled as a network (Figure 2.5). This frame of thinking closely resembles that of computer architecture, in which subsystems have particular states that change based on the states of other subsystems (Neumann, 1993). In a computer, the rules of changing states and the rules of interaction between states are also determined by stimuli or the states of other subsystems. While the brain is not such a computer, its function may yet be based on subsystem representations and rules for how they interact and transition from one representation to another. Moreover, recent developments in artificial intelligence (AI) highlight the importance of understanding how neural networks form representations. For example, reinforcement learning, a subfield in AI as well as in neuroscience (Neftci et al., 2019), uses neural networks to learn abstract representations of the environment to play complex games like Go (Silver et al., 2016) and DOTA 2 (OpenAI et al., 2019) without human supervision. In this review, we have aimed to outline the evidence for and methods to study the dynamic representations on neural networks that underlie cognition and behaviors.

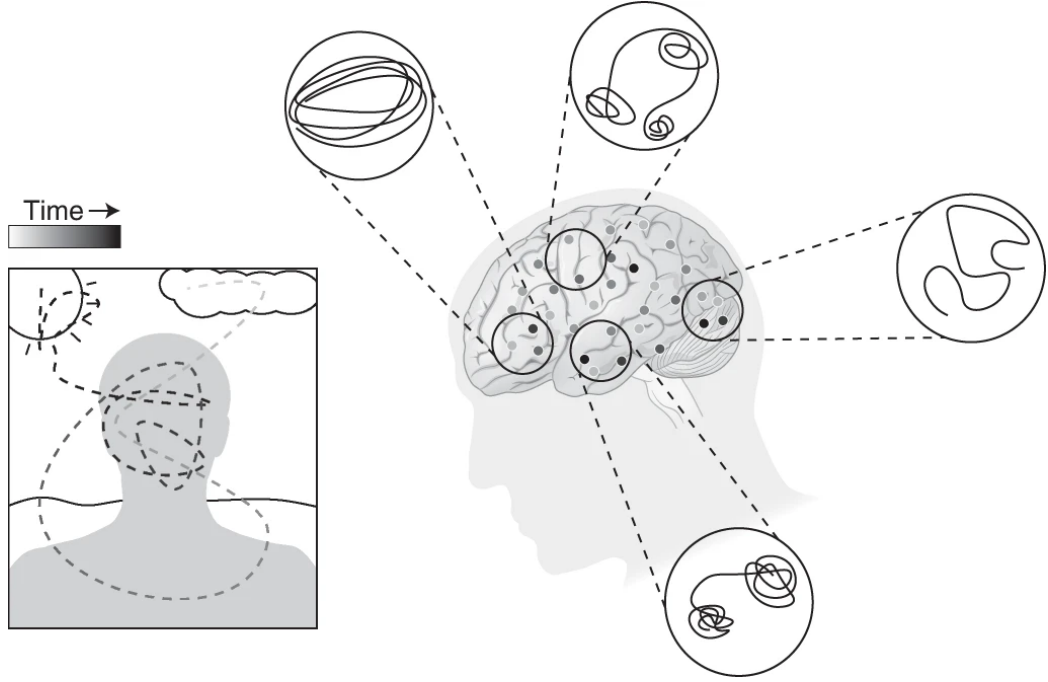


Figure 2.5: **Dynamic representations as trajectories in neural state space.** As we view our environment (framed box), the representations in various neural subsystems (circled) can evolve differently (insets inside circles) and interact with others to perform perception, cognition and behaviors.

## 2.7. Conclusion

Organisms interact with an ever-changing environment. Recent studies extend our understanding of how organisms model the world by investigating how representations change over time and how they are transmitted across neurons and brain regions. However, further work is required to understand the dynamics of representations; we propose that the integration between neural representations and network models can accelerate this progress. Thus, we build a framework for dynamic representations that describes (i) intraregional dynamics and (ii) the interactions between intra- and inter-regional dynamics. We organize evidence from the literature supporting these mechanisms. Finally, we review important frontiers in understanding the dynamic representations that support cognition. Altogether, the framework of dynamic representations begins to reveal how the dynamics of neural systems support cognition and may further elucidate the crucial crossover from matter to mind.

## REFERENCES

- Anzellotti, Stefano, Alfonso Caramazza, and Rebecca Saxe (2017). “Multivariate pattern dependence.” In: *PLOS Computational Biology* 13.11. Ed. by Saad Jbabdi, e1005799. DOI: 10.1371/journal.pcbi.1005799.
- Arana, Sophie et al. (2020). “Sensory Modality-Independent Activation of the Brain Network for Language.” In: *The Journal of Neuroscience* 40.14, pp. 2914–2924. DOI: 10.1523/JNEUROSCI.2271-19.2020.
- Aston-Jones, Gary and Jonathan D. Cohen (2005). “An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance.” In: *Annual Review of Neuroscience* 28.1, pp. 403–450. DOI: 10.1146/annurev.neuro.28.061604.135709.
- Avena-Koenigsberger, Andrea, Bratislav Misic, and Olaf Sporns (2018). “Communication dynamics in complex brain networks.” In: *Nature Reviews Neuroscience* 19.1, pp. 17–33. DOI: 10.1038/nrn.2017.149.
- Badre, David and Theresa M. Desrochers (2019). “Hierarchical cognitive control and the frontal lobes.” In: *Handbook of Clinical Neurology*. Vol. 163. Elsevier, pp. 165–177. ISBN: 978-0-12-804281-6. DOI: 10.1016/B978-0-12-804281-6.00009-4.
- Bansal, Kanika, Johan Nakuci, and Sarah Feldt Muldoon (2018). “Personalized brain network models for assessing structure–function relationships.” In: *Current Opinion in Neurobiology* 52, pp. 42–47. DOI: 10.1016/j.conb.2018.04.014.
- Barbas, Helen and Basilis Zikopoulos (2007). “The Prefrontal Cortex and Flexible Behavior.” In: *The Neuroscientist* 13.5, pp. 532–545. DOI: 10.1177/1073858407301369.
- Bartos, Marlene, Yair Manor, et al. (1999). “Coordination of Fast and Slow Rhythmic Neuronal Circuits.” In: *The Journal of Neuroscience* 19.15, pp. 6650–6660. DOI: 10.1523/JNEUROSCI.19-15-06650.1999.
- Bartos, Marlene, Imre Vida, and Peter Jonas (2007). “Synaptic mechanisms of synchronized gamma oscillations in inhibitory interneuron networks.” In: *Nature Reviews Neuroscience* 8.1, pp. 45–56. DOI: 10.1038/nrn2044.

- Bassett, Danielle S., Nicholas F. Wymbs, et al. (2011). “Dynamic reconfiguration of human brain networks during learning.” In: *Proceedings of the National Academy of Sciences* 108.18, pp. 7641–7646. DOI: 10.1073/pnas.1018985108.
- Bassett, Danielle S., Perry Zurn, and Joshua I. Gold (2018). “On the nature and use of models in network neuroscience.” In: *Nature Reviews Neuroscience* 19.9, pp. 566–578. DOI: 10.1038/s41583-018-0038-8.
- Bassett, Danielle S and Olaf Sporns (2017). “Network neuroscience.” In: *Nature Neuroscience* 20.3, pp. 353–364. DOI: 10.1038/nn.4502.
- Bassett, Danielle Smith and Ed Bullmore (2006). “Small-World Brain Networks.” In: *The Neuroscientist* 12.6, pp. 512–523. DOI: 10.1177/1073858406293182.
- Battaglia, Peter W. et al. (2012). “Computational Models of Intuitive Physics.” In: *Cognitive Science* 34.
- Becker, Cassiano O, Danielle S Bassett, and Victor M Preciado (2018). “Large-scale dynamic modeling of task-fMRI signals via subspace system identification.” In: *Journal of Neural Engineering* 15.6, p. 066016. DOI: 10.1088/1741-2552/aad8c7.
- Beggs, J. M. (2004). “Neuronal Avalanches Are Diverse and Precise Activity Patterns That Are Stable for Many Hours in Cortical Slice Cultures.” In: *Journal of Neuroscience* 24.22, pp. 5216–5229. DOI: 10.1523/JNEUROSCI.0540-04.2004.
- Bertolero, Maxwell A., B. T. Thomas Yeo, and Mark D’Esposito (2015). “The modular and integrative functional architecture of the human brain.” In: *Proceedings of the National Academy of Sciences* 112.49. DOI: 10.1073/pnas.1510619112.
- Bettinardi, R. G. et al. (2017). “How structure sculpts function: Unveiling the contribution of anatomical connectivity to the brain’s spontaneous correlation structure.” In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 27.4, p. 047409. DOI: 10.1063/1.4980099.
- Borg-Graham, Lyle J., Cyril Monier, and Yves Frégnac (1998). “Visual input evokes transient and strong shunting inhibition in visual cortical neurons.” In: *Nature* 393.6683, pp. 369–373. DOI: 10.1038/30735.

- Breakspear, Michael (2017). “Dynamic models of large-scale brain activity.” In: *Nature Neuroscience* 20.3, pp. 340–352. DOI: 10.1038/nn.4497.
- Brunec, Iva K., Morris Moscovitch, and Morgan D. Barense (2018). “Boundaries Shape Cognitive Representations of Spaces and Events.” In: *Trends in Cognitive Sciences* 22.7, pp. 637–650. DOI: 10.1016/j.tics.2018.03.013.
- Bullmore, Edward and Petra Vértes (2013). “From Lichtheim to Rich Club: Brain Networks and Psychiatry.” In: *JAMA Psychiatry* 70.8, p. 780. DOI: 10.1001/jamapsychiatry.2013.212.
- Buzsáki, György and Andreas Draguhn (2004). “Neuronal Oscillations in Cortical Networks.” In: *Science* 304.5679, pp. 1926–1929. DOI: 10.1126/science.1099745.
- Chang, Catie et al. (2016). “Tracking brain arousal fluctuations with fMRI.” In: *Proceedings of the National Academy of Sciences* 113.16, pp. 4518–4523. DOI: 10.1073/pnas.1520613113.
- Chaudhuri, Rishidev et al. (2019). “The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep.” In: *Nature Neuroscience* 22.9, pp. 1512–1520. DOI: 10.1038/s41593-019-0460-x.
- Chun, Marvin M., Julie D. Golomb, and Nicholas B. Turk-Browne (2011). “A Taxonomy of External and Internal Attention.” In: *Annual Review of Psychology* 62.1, pp. 73–101. DOI: 10.1146/annurev.psych.093008.100427.
- Clancy, Kelly B., Ivana Orsolic, and Thomas D. Mrsic-Flogel (2019). “Locomotion-dependent remapping of distributed cortical networks.” In: *Nature Neuroscience* 22.5, pp. 778–786. DOI: 10.1038/s41593-019-0357-8.
- Cohen, Marlene R. and William T. Newsome (2008). “Context-Dependent Changes in Functional Circuitry in Visual Area MT.” In: *Neuron* 60.1, pp. 162–173. DOI: 10.1016/j.neuron.2008.08.007.
- Coleman, Melissa J., Pierre Meyrand, and Michael P. Nusbaum (1995). “A switch between two modes of synaptic transmission mediated by presynaptic inhibition.” In: *Nature* 378.6556, pp. 502–505. DOI: 10.1038/378502a0.

- Constantinescu, Alexandra O., Jill X. O'Reilly, and Timothy E. J. Behrens (2016). "Organizing conceptual knowledge in humans with a gridlike code." In: *Science* 352.6292, pp. 1464–1468. DOI: 10.1126/science.aaf0941.
- Cools, Roshan et al. (2019). "Dopamine and the motivation of cognitive control." In: *Handbook of Clinical Neurology*. Vol. 163. Elsevier, pp. 123–143. ISBN: 978-0-12-804281-6. DOI: 10.1016/B978-0-12-804281-6.00007-0.
- Cornblath, Eli J. et al. (2020). "Temporal sequences of brain activity at rest are constrained by white matter structure and modulated by cognitive demands." In: *Communications Biology* 3.1, p. 261. DOI: 10.1038/s42003-020-0961-x.
- Coutanche, Marc N. and Sharon L. Thompson-Schill (2013). "Informational connectivity: identifying synchronized discriminability of multi-voxel patterns across the brain." In: *Frontiers in Human Neuroscience* 7. DOI: 10.3389/fnhum.2013.00015.
- Cummings, Kirstie A. and Roger L. Clem (2020). "Prefrontal somatostatin interneurons encode fear memory." In: *Nature Neuroscience* 23.1, pp. 61–74. DOI: 10.1038/s41593-019-0552-7.
- Curry, Justin (2014). "Sheaves, Cosheaves and Applications." In: *arXiv:1303.3255 [math]*.
- Danjo, Teruko, Taro Toyozumi, and Shigeyoshi Fujisawa (2018). "Spatial representations of self and other in the hippocampus." In: *Science* 359.6372, pp. 213–218. DOI: 10.1126/science.aao3898.
- DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust (2012). "How Does the Brain Solve Visual Object Recognition?" In: *Neuron* 73.3, pp. 415–434. DOI: 10.1016/j.neuron.2012.01.010.
- Dragoi, George and György Buzsáki (2006). "Temporal Encoding of Place Sequences by Hippocampal Cell Assemblies." In: *Neuron* 50.1, pp. 145–157. DOI: 10.1016/j.neuron.2006.02.023.
- Ekstrom, Arne D. et al. (2003). "Cellular networks underlying human spatial navigation." In: *Nature* 425.6954, pp. 184–188. DOI: 10.1038/nature01964.



- Eldar, Eran, Jonathan D Cohen, and Yael Niv (2013). “The effects of neural gain on attention and learning.” In: *Nature Neuroscience* 16.8, pp. 1146–1153. DOI: 10.1038/nn.3428.
- Eliasmith, Chris et al. (2012). “A Large-Scale Model of the Functioning Brain.” In: *Science* 338.6111, pp. 1202–1205. DOI: 10.1126/science.1225266.
- Epstein, Russell A et al. (2017). “The cognitive map in humans: spatial navigation and beyond.” In: *Nature Neuroscience* 20.11, pp. 1504–1513. DOI: 10.1038/nn.4656.
- Eschbach, Claire et al. (2020). “Recurrent architecture for adaptive regulation of learning in the insect brain.” In: *Nature Neuroscience* 23.4, pp. 544–555. DOI: 10.1038/s41593-020-0607-9.
- Faber, Samantha P. et al. (2019). “Computation is concentrated in rich clubs of local cortical networks.” In: *Network Neuroscience* 3.2, pp. 384–404. DOI: 10.1162/netn\_a\_00069.
- Fox, Michael D. (2018). “Mapping Symptoms to Brain Networks with the Human Connectome.” In: *New England Journal of Medicine* 379.23, pp. 2237–2245. DOI: 10.1056/NEJMr1706158.
- Fox, Michael D. et al. (2005). “The human brain is intrinsically organized into dynamic, anticorrelated functional networks.” In: *Proceedings of the National Academy of Sciences* 102.27, pp. 9673–9678. DOI: 10.1073/pnas.0504136102.
- Friston, K.J., L. Harrison, and W. Penny (2003). “Dynamic causal modelling.” In: *NeuroImage* 19.4, pp. 1273–1302. DOI: 10.1016/S1053-8119(03)00202-7.
- Friston, Karl J. (2011). “Functional and Effective Connectivity: A Review.” In: *Brain Connectivity* 1.1, pp. 13–36. DOI: 10.1089/brain.2011.0008.
- Gallego, Juan A. et al. (2018). “Cortical population activity within a preserved neural manifold underlies multiple motor behaviors.” In: *Nature Communications* 9.1, p. 4233. DOI: 10.1038/s41467-018-06560-z.
- Greene, Clint et al. (2019). “Finding maximally disconnected subnetworks with shortest path tractography.” In: *NeuroImage: Clinical* 23, p. 101903. DOI: 10.1016/j.nicl.2019.101903.

- Gu, Shi, Matthew Cieslak, et al. (2018). “The Energy Landscape of Neurophysiological Activity Implicit in Brain Network Structure.” In: *Scientific Reports* 8.1, p. 2507. DOI: 10.1038/s41598-018-20123-8.
- Gu, Shi, Fabio Pasqualetti, et al. (2015). “Controllability of structural brain networks.” In: *Nature Communications* 6.1, p. 8414. DOI: 10.1038/ncomms9414.
- Haider, Bilal and David A. McCormick (2009). “Rapid Neocortical Dynamics: Cellular and Network Mechanisms.” In: *Neuron* 62.2, pp. 171–189. DOI: 10.1016/j.neuron.2009.04.008.
- Hassabis, D., D. Kumaran, and E. A. Maguire (2007). “Using Imagination to Understand the Neural Basis of Episodic Memory.” In: *Journal of Neuroscience* 27.52, pp. 14365–14374. DOI: 10.1523/JNEUROSCI.4549-07.2007.
- Hatsopoulos, Nicholas G. and John P. Donoghue (2009). “The Science of Neural Interface Systems.” In: *Annual Review of Neuroscience* 32.1, pp. 249–266. DOI: 10.1146/annurev.neuro.051508.135241.
- Henzinger, T.A. (1996). “The theory of hybrid automata.” In: *Proceedings 11th Annual IEEE Symposium on Logic in Computer Science*. New Brunswick, NJ, USA: IEEE Comput. Soc. Press, pp. 278–292. ISBN: 978-0-8186-7463-1. DOI: 10.1109/LICS.1996.561342.
- Hermundstad, Ann M. et al. (2013). “Structural foundations of resting-state and task-based functional connectivity in the human brain.” In: *Proceedings of the National Academy of Sciences* 110.15, pp. 6169–6174. DOI: 10.1073/pnas.1219562110.
- Heuvel, M. P. van den and O. Sporns (2011). “Rich-Club Organization of the Human Connectome.” In: *Journal of Neuroscience* 31.44, pp. 15775–15786. DOI: 10.1523/JNEUROSCI.3539-11.2011.
- Heuvel, M.P. van den, C.J. Stam, et al. (2008). “Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain.” In: *NeuroImage* 43.3, pp. 528–539. DOI: 10.1016/j.neuroimage.2008.08.010.
- Hopfield, John J. and David W. Tank (1986). “Computing with Neural Circuits: A Model.” In: *Science* 233.4764, pp. 625–633. DOI: 10.1126/science.3755256.

- Ito, Shinya, Michael E. Hansen, et al. (2011). “Extending Transfer Entropy Improves Identification of Effective Connectivity in a Spiking Cortical Network Model.” In: *PLoS ONE* 6.11. Ed. by Michal Zochowski, e27431. DOI: 10.1371/journal.pone.0027431.
- Ito, Takuya, Luke Hearne, et al. (2020). “Discovering the Computational Relevance of Brain Network Organization.” In: *Trends in Cognitive Sciences* 24.1, pp. 25–38. DOI: 10.1016/j.tics.2019.10.005.
- Johansen-Berg, Heidi (2013). “Human connectomics — What will the future demand?” In: *NeuroImage* 80, pp. 541–544. DOI: 10.1016/j.neuroimage.2013.05.082.
- Joo, Hannah R. and Loren M. Frank (2018). “The hippocampal sharp wave–ripple in memory retrieval for immediate use and consolidation.” In: *Nature Reviews Neuroscience* 19.12, pp. 744–757. DOI: 10.1038/s41583-018-0077-1.
- Ju, Harang et al. (2020). “Network structure of cascading neural systems predicts stimulus propagation and recovery.” In: *Journal of Neural Engineering* 17.5, p. 056045. DOI: 10.1088/1741-2552/abbff1.
- Kailath, Thomas (1980). *Linear systems*. Prentice-Hall information and system science series. Englewood Cliffs, N.J: Prentice-Hall. ISBN: 978-0-13-536961-6.
- Katz, Paul S. and William N. Frost (1996). “Intrinsic neuromodulation: altering neuronal circuits from within.” In: *Trends in Neurosciences* 19.2, pp. 54–61. DOI: 10.1016/0166-2236(96)89621-4.
- Kiltner, Konstantina and H. Henrik Ehrsson (2020). “Functional Connectivity between the Cerebellum and Somatosensory Areas Implements the Attenuation of Self-Generated Touch.” In: *The Journal of Neuroscience* 40.4, pp. 894–906. DOI: 10.1523/JNEUROSCI.1732-19.2019.
- Kim, Jason Z. et al. (2018). “Role of graph architecture in controlling dynamical networks with applications to neural systems.” In: *Nature Physics* 14.1, pp. 91–98. DOI: 10.1038/nphys4268.

- Kirst, Christoph, Marc Timme, and Demian Battaglia (2016). “Dynamic information routing in complex networks.” In: *Nature Communications* 7.1, p. 11061. DOI: 10.1038/ncomms11061.
- Krabbe, Sabine et al. (2019). “Adaptive disinhibitory gating by VIP interneurons permits associative learning.” In: *Nature Neuroscience* 22.11, pp. 1834–1843. DOI: 10.1038/s41593-019-0508-y.
- Kriegeskorte, Nikolaus (2008). “Representational similarity analysis – connecting the branches of systems neuroscience.” In: *Frontiers in Systems Neuroscience*. DOI: 10.3389/neuro.06.004.2008.
- Kumaran, Dharshan, Demis Hassabis, and James L. McClelland (2016). “What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated.” In: *Trends in Cognitive Sciences* 20.7, pp. 512–534. DOI: 10.1016/j.tics.2016.05.004.
- Lavie, Nilli et al. (2004). “Load Theory of Selective Attention and Cognitive Control.” In: *Journal of Experimental Psychology: General* 133.3, pp. 339–354. DOI: 10.1037/0096-3445.133.3.339.
- Levy, Dana Rubi et al. (2019). “Dynamics of social representation in the mouse prefrontal cortex.” In: *Nature Neuroscience* 22.12, pp. 2013–2022. DOI: 10.1038/s41593-019-0531-z.
- Lisman, John E. and Marco A. P. Idiart (1995). “Storage of  $7 \pm 2$  Short-Term Memories in Oscillatory Subcycles.” In: *Science* 267.5203, pp. 1512–1515. DOI: 10.1126/science.7878473.
- Liu, Yang-Yu, Jean-Jacques Slotine, and Albert-László Barabási (2011). “Controllability of complex networks.” In: *Nature* 473.7346, pp. 167–173. DOI: 10.1038/nature10011.
- Mahmoudi, Abdelhak et al. (2012). “Multivoxel Pattern Analysis for fMRI Data: A Review.” In: *Computational and Mathematical Methods in Medicine* 2012, pp. 1–14. DOI: 10.1155/2012/961257.
- Marr, David and H K Nishihara (1978). “Representation and recognition of the spatial organization of three-dimensional shapes.” In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 200.1140, pp. 269–294. DOI: 10.1098/rspb.1978.0020.

- McIntosh, Anthony R. and Bratislav Mišić (2013). “Multivariate Statistical Analyses for Neuroimaging Data.” In: *Annual Review of Psychology* 64.1, pp. 499–525. DOI: 10.1146/annurev-psych-113011-143804.
- Mišić, Bratislav et al. (2015). “Cooperative and Competitive Spreading Dynamics on the Human Connectome.” In: *Neuron* 86.6, pp. 1518–1529. DOI: 10.1016/j.neuron.2015.05.035.
- Mobbs, Dean et al. (2020). “Space, Time, and Fear: Survival Computations along Defensive Circuits.” In: *Trends in Cognitive Sciences* 24.3, pp. 228–241. DOI: 10.1016/j.tics.2019.12.016.
- Musall, Simon et al. (2019). “Single-trial neural dynamics are dominated by richly varied movements.” In: *Nature Neuroscience* 22.10, pp. 1677–1686. DOI: 10.1038/s41593-019-0502-4.
- Neftci, Emre O. and Bruno B. Averbeck (2019). “Reinforcement learning in artificial and biological systems.” In: *Nature Machine Intelligence* 1.3, pp. 133–143. DOI: 10.1038/s42256-019-0025-4.
- Neumaier, Arnold and Tapio Schneider (2001). “Estimation of parameters and eigenmodes of multivariate autoregressive models.” In: *ACM Transactions on Mathematical Software* 27.1, pp. 27–57. DOI: 10.1145/382043.382304.
- Neumann, J. von (1993). “First draft of a report on the EDVAC.” In: *IEEE Annals of the History of Computing* 15.4, pp. 27–75. DOI: 10.1109/85.238389.
- Nozari, Erfan and Jorge Cortes (2021). “Hierarchical Selective Recruitment in Linear-Threshold Brain Networks Part II: Multilayer Dynamics and Top-Down Recruitment.” In: *IEEE Transactions on Automatic Control* 66.3, pp. 965–980. DOI: 10.1109/TAC.2020.2997854.
- Nusbaum, Michael P. and Mark P. Beenhakker (2002). “A small-systems approach to motor pattern generation.” In: *Nature* 417.6886, pp. 343–350. DOI: 10.1038/417343a.
- OpenAI et al. (2019). “Dota 2 with Large Scale Deep Reinforcement Learning.” In: *arXiv:1912.06680 [cs, stat]*.
- Palmigiano, Agostina et al. (2017). “Flexible information routing by transient synchrony.” In: *Nature Neuroscience* 20.7, pp. 1014–1022. DOI: 10.1038/nn.4569.

- Parthasarathy, Aishwarya et al. (2017). “Mixed selectivity morphs population codes in pre-frontal cortex.” In: *Nature Neuroscience* 20.12, pp. 1770–1779. DOI: 10.1038/s41593-017-0003-2.
- Popov, Tzvetan et al. (2018). “Time Course of Brain Network Reconfiguration Supporting Inhibitory Control.” In: *The Journal of Neuroscience* 38.18, pp. 4348–4356. DOI: 10.1523/JNEUROSCI.2639-17.2018.
- Richmond, Lauren L. and Jeffrey M. Zacks (2017). “Constructing Experience: Event Models from Perception to Action.” In: *Trends in Cognitive Sciences* 21.12, pp. 962–980. DOI: 10.1016/j.tics.2017.08.005.
- Rigotti, Mattia et al. (2013). “The importance of mixed selectivity in complex cognitive tasks.” In: *Nature* 497.7451, pp. 585–590. DOI: 10.1038/nature12160.
- Roberts, James A. et al. (2019). “Metastable brain waves.” In: *Nature Communications* 10.1, p. 1056. DOI: 10.1038/s41467-019-08999-0.
- Rombach, M. Puck et al. (2014). “Core-Periphery Structure in Networks.” In: *SIAM Journal on Applied Mathematics* 74.1, pp. 167–190. DOI: 10.1137/120881683.
- Saxena, Shreya and John P Cunningham (2019). “Towards the neural population doctrine.” In: *Current Opinion in Neurobiology* 55, pp. 103–111. DOI: 10.1016/j.conb.2019.02.002.
- Schapiro, Anna C et al. (2013). “Neural representations of events arise from temporal community structure.” In: *Nature Neuroscience* 16.4, pp. 486–492. DOI: 10.1038/nn.3331.
- Shannon, Claude Elwood and Warren Weaver (1998). *The mathematical theory of communication*. Urbana: University of Illinois Press. ISBN: 978-0-252-72546-3.
- Shew, Woodrow L. et al. (2015). “Adaptation to sensory input tunes visual cortex to criticality.” In: *Nature Physics* 11.8, pp. 659–663. DOI: 10.1038/nphys3370.
- Shine, James M. et al. (2019). “Human cognition involves the dynamic integration of neural activity and neuromodulatory systems.” In: *Nature Neuroscience* 22.2, pp. 289–296. DOI: 10.1038/s41593-018-0312-0.
- Silver, David et al. (2016). “Mastering the game of Go with deep neural networks and tree search.” In: *Nature* 529.7587, pp. 484–489. DOI: 10.1038/nature16961.

- Sohn, Hansem et al. (2019). “Bayesian Computation through Cortical Latent Dynamics.” In: *Neuron* 103.5, 934–947.e5. DOI: 10.1016/j.neuron.2019.06.012.
- Spelke, Elizabeth S. and Katherine D. Kinzler (2007). “Core knowledge.” In: *Developmental Science* 10.1, pp. 89–96. DOI: 10.1111/j.1467-7687.2007.00569.x.
- Sporns, Olaf and Richard F. Betzel (2016). “Modular Brain Networks.” In: *Annual Review of Psychology* 67.1, pp. 613–640. DOI: 10.1146/annurev-psych-122414-033634.
- Stiso, Jennifer et al. (2019). “White Matter Network Architecture Guides Direct Electrical Stimulation through Optimal State Transitions.” In: *Cell Reports* 28.10, 2554–2566.e7. DOI: 10.1016/j.celrep.2019.08.008.
- Stringer, Carsen, Marius Pachitariu, Nicholas A Steinmetz, Michael Okun, et al. (2016). “Inhibitory control of correlated intrinsic variability in cortical networks.” In: *eLife* 5, e19695. DOI: 10.7554/eLife.19695.
- Stringer, Carsen, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, et al. (2019). “High-dimensional geometry of population responses in visual cortex.” In: *Nature* 571.7765, pp. 361–365. DOI: 10.1038/s41586-019-1346-5.
- Stringer, Carsen, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, et al. (2019). “Spontaneous behaviors drive multidimensional, brainwide activity.” In: *Science* 364.6437, eaav7893. DOI: 10.1126/science.aav7893.
- Summerfield, Christopher et al. (2006). “Predictive Codes for Forthcoming Perception in the Frontal Cortex.” In: *Science* 314.5803, pp. 1311–1314. DOI: 10.1126/science.1132028.
- Taghia, Jalil et al. (2018). “Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition.” In: *Nature Communications* 9.1, p. 2505. DOI: 10.1038/s41467-018-04723-6.
- Tang, Evelyn et al. (2019). “Effective learning is accompanied by high-dimensional and efficient representations of neural activity.” In: *Nature Neuroscience* 22.6, pp. 1000–1009. DOI: 10.1038/s41593-019-0400-9.
- Tavares, Rita Morais et al. (2015). “A Map for Social Navigation in the Human Brain.” In: *Neuron* 87.1, pp. 231–243. DOI: 10.1016/j.neuron.2015.06.011.

- Téglás, Ernő et al. (2011). “Pure Reasoning in 12-Month-Old Infants as Probabilistic Inference.” In: *Science* 332.6033, pp. 1054–1059. DOI: 10.1126/science.1196404.
- Tschida, Katherine et al. (2019). “A Specialized Neural Circuit Gates Social Vocalizations in the Mouse.” In: *Neuron* 103.3, 459–472.e4. DOI: 10.1016/j.neuron.2019.05.025.
- Vázquez-Rodríguez, Bertha et al. (2019). “Gradients of structure–function tethering across neocortex.” In: *Proceedings of the National Academy of Sciences* 116.42, pp. 21219–21227. DOI: 10.1073/pnas.1903403116.
- Weber, Alison I and Adrienne L Fairhall (2019). “The role of adaptation in neural coding.” In: *Current Opinion in Neurobiology* 58, pp. 135–140. DOI: 10.1016/j.conb.2019.09.013.
- Wibral, Michael et al. (2017). “Quantifying Information Modification in Developing Neural Networks via Partial Information Decomposition.” In: *Entropy* 19.9, p. 494. DOI: 10.3390/e19090494.
- Wilting, J and V Priesemann (2019). “25 years of criticality in neuroscience — established results, open controversies, novel concepts.” In: *Current Opinion in Neurobiology* 58, pp. 105–111. DOI: 10.1016/j.conb.2019.08.002.
- Wolf, Fred et al. (2014). “Dynamical models of cortical circuits.” In: *Current Opinion in Neurobiology* 25, pp. 228–236. DOI: 10.1016/j.conb.2014.01.017.
- Yan, Gang et al. (2017). “Network control principles predict neuron function in the *Caenorhabditis elegans* connectome.” In: *Nature* 550.7677, pp. 519–523. DOI: 10.1038/nature24056.
- Yang, Guangyu Robert et al. (2019). “Task representations in neural networks trained to perform many cognitive tasks.” In: *Nature Neuroscience* 22.2, pp. 297–306. DOI: 10.1038/s41593-018-0310-2.



## CHAPTER 3

### MEMORY IN CASCADING NEURAL NETWORKS

*This chapter contains work from Ju, H., Kim, J.Z., Beggs, J.M., and Bassett, D.S. (2020). “Network structure of cascading neural systems predicts stimulus propagation and recovery.” Journal of Neural Engineering 17, 056045.*

#### 3.1. Abstract

Many neural systems display spontaneous, spatiotemporal patterns of neural activity that are crucial for information processing. While these cascading patterns presumably arise from the underlying network of synaptic connections between neurons, the precise contribution of the network’s local and global connectivity to these patterns and information processing remains largely unknown. Here, we demonstrate how network structure supports information processing through network dynamics in empirical and simulated spiking neurons using mathematical tools from linear systems theory, network control theory, and information theory. In particular, we show that activity, and the information that it contains, travels through cycles in real and simulated networks. Broadly, our results demonstrate how cascading neural networks could contribute to cognitive faculties that require lasting activation of neuronal patterns, such as working memory or attention.

#### 3.2. Introduction

A central question in neuroscience is how connections between neurons determine patterns of neurophysiological activity that support organism function. Networks of neurons receive incoming stimuli and perform computations to shape cognition and behavior, such as the visual recognition of faces in regulating social behavior (Adolphs, 2003). While many studies laud the ultimate goal of determining how the brain’s network structure supports information processing (Watts et al., 1998; Honey et al., 2007), it remains challenging to

empirically study the direct interactions between neural dynamics, connectivity, and computation. Indeed, neural connections and their underlying computational function have often been inferred through neural dynamics, and formal studies probing mechanistic relations among the three components have remained largely theoretical (Hopfield, 1982; Ben-Yishai et al., 1995; Wang, 2002).

One characteristic empirical feature of many systems is cascading dynamics, in which neurons display spontaneous bursts of activity. While these bursts may seem arbitrary, they actually comprise stochastic cascades that follow spatiotemporal patterns of activity (Haldeman et al., 2005). These cortical cascading dynamics have been well-characterized in the empirical literature using a range of methods *in vitro* (Beggs and Plenz, 2003; Beggs, 2004), *in vivo* (Gireesh et al., 2008; Petermann et al., 2009; Hahn et al., 2010; Shriki et al., 2013; Bellay et al., 2015; Ponce-Alvarez et al., 2018), and *ex vivo* (Shew, Clawson, et al., 2015) in a variety of organisms, including humans. In a complementary line of theoretical work, these neural systems have been hypothesized to operate within a regime that maximizes information transmission (Beggs and Plenz, 2003; Shew, Yang, Yu, et al., 2011), information storage (Haldeman et al., 2005), computational power (Bertschinger et al., 2004), and dynamic range (Kinouchi et al., 2006; Shew, Yang, Petermann, et al., 2009; Larremore, Shew, Ott, et al., 2011). However, often left implicit in these analyses is the structure of the networks underlying such dynamics and how the structure may constrain those dynamics. Relatively recent empirical data show evidence of specific patterns of cortical connectivity. Cortical neurons are often strongly, bidirectionally connected to each other (Wang, Markram, et al., 2006; Lefort et al., 2009; Ko et al., 2011), and form higher-order network motifs in clusters of neurons (Markram, 1997; Song, Sjöström, et al., 2005; Perin et al., 2011), which in turn group into communities of neurons. By forming cyclical network motifs, neurons can temporally extend activity, thereby supporting short-term information storage (Rodriguez et al., 2001; Fiete et al., 2010; Daie et al., 2015; Brunel, 2016) and computation (Shimono et al., 2014; Nigam et al., 2016; Faber et al., 2019). These features of network structure have yet to be linked to cascading dynamics and computations that are supported by those same circuits.

Here, we address the gap in knowledge between network structure, cascading dynamics, and information retrieval through a series of analytical and numerical analyses on simulated and empirical data. We first show that network structure constrains system memory through sustained activity by framing spike propagation as state transitions in a Markov chain (Howard, 1971). We then apply linear systems theory to predict distributions of cascade duration in the stochastic dynamics of simulated and empirical spiking neural networks. We find that cascades flow through cycles in the underlying network, which have been widely observed in experiments, to contribute to long tails in distributions of cascade duration. Finally, we use mutual information to probe the relations among network structure, cascade duration, and the information maintained in a network in 4 commonly studied generative graph models. Moreover, our method can accommodate networks that are both non-critical (Touboul et al., 2010; Friedman et al., 2012; Priesemann et al., 2014; Touboul et al., 2017) and critical, and that show avalanche behavior, characterized by power-law distributions of cascade *size* (i.e., the number of neurons that spike in a cascade) and *duration*. Collectively, our findings show that the network topology reported extensively in the empirical literature can produce complex cascading dynamics through which a network can support the lasting activation of a cluster of neurons, which in turn allows for the discrimination of stimulus patterns implicated in working memory (Goldman-Rakic, 1995; Durstewitz et al., 2000; Eriksson et al., 2015).

### 3.3. Mathematical Framework

#### 3.3.1. Network formulation

We begin with the stipulation of a network as well as a dynamical process that occurs atop the network. We formalize the notion of a network as a directed graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  in which neurons are represented as nodes  $\mathcal{V} = \{1, \dots, n\}$  and neuron-to-neuron connections are represented as edges  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$  (Figure 3.1A). The weighted and directed adjacency matrix  $A = [a_{ij}]$  thus encodes the edge weights from neuron  $j$  to neuron  $i$  (Figure 3.1B).

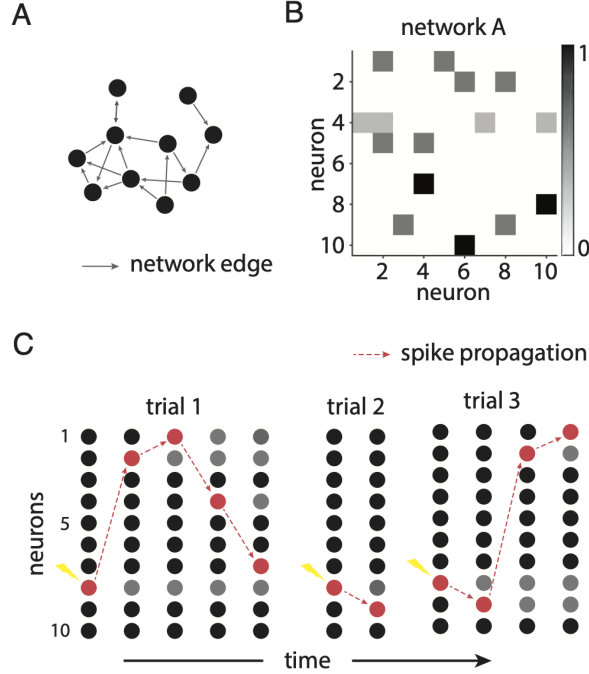


Figure 3.1: **A linear dynamical system accurately estimates the average spiking of neurons in a stochastic model.** **A** An example network represented as an adjacency matrix  $A$ . **B** A Markov chain of network states can accurately predict the fraction of active cascades at time  $t$ . In  $10^4$  trials of stimulating neuron 8 in the network in panel b, the root-mean-square error between the state-space prediction and the stochastic model is  $1.2 \times 10^{-4}$ . **C** Examples of simulations of cascades generated by stimulating neuron 8 in the network in panel b.

### 3.3.2. Stochastic McCulloch-Pitts neuron

To model neuronal cascades, we next stipulate a stochastic, discrete-time version of the McCulloch-Pitts neuron (McCulloch et al., 1990). In the McCulloch-Pitts model, a neuron  $i$  receives inputs scaled by the weights of the edges and sums the scaled inputs,  $\mathbf{a}_i \cdot \mathbf{y}$ , to produce an output spike  $y_i(t)$  via an activation function (Figure 3.1C). Here, the activation function is a random Bernoulli process, where probability  $p$  is the sum of the scaled inputs. The sum of the scaled inputs  $\mathbf{a}_i \cdot \mathbf{y}$  is bound by 0 and 1 such that  $p = \min(1, \max(0, \mathbf{a}_i \cdot \mathbf{y}))$ . The network starts at some non-random initial state  $\mathbf{y}(0)$ , which can also be interpreted as a stimulus received at  $t = 0$ . The state of an  $n$ -neuron network is a binary vector  $\mathbf{y}(t) \in \{0, 1\}^n$

such that each element indicates whether a neuron fired at time  $t$  and evolves as

$$y_i(t) \sim B\left(\min(1, \max(0, \mathbf{a}_i \cdot \mathbf{y}))\right), \quad (3.1)$$

where  $B(p)$  is a Bernoulli process with probability  $p$ , and  $\mathbf{a}_i$  is the  $i^{th}$  row vector of  $A$ .

### 3.3.3. Markov chain formulation

The model that we consider can be represented as a Markov chain with states  $\mathbf{s}^i \in \{0, 1\}^n$  representing all possible patterns of spikes in the network, and with state  $\mathbf{s}^1 = \mathbf{0}$  representing the zero state. The column vector  $\mathbf{p}(t) = [p_1(t); \dots; p_{2^n}(t)] = [P(\mathbf{y}(t) = \mathbf{s}^1); \dots; P(\mathbf{y}(t) = \mathbf{s}^{2^n})]$  represents the probability that the network exists in any state  $\mathbf{s}^i$  at time  $t$ . The transition matrix  $T$  governs

$$\mathbf{p}(t) = T\mathbf{p}(t-1) = T^t\mathbf{p}(0), \quad (3.2)$$

where each entry  $T = [T_{lk}] = P([\mathbf{y}(t) = \mathbf{s}^l] | [\mathbf{y}(t-1) = \mathbf{s}^k])$  represents the transition probability from state  $k$  to state  $l$ . See S1 Methods for details regarding the computation of the matrix  $T$  and S1 Result for numerical validation. With the transition matrix, we can compute the fundamental matrix of the Markov chain (Howard, 1971).

### 3.3.4. Estimation as a linear dynamical system

Because computing a Markov chain is intractable for large network sizes, we instead estimate the process stated in Equation 3.1 using a linear dynamical system with the same parameters  $A$ . Specifically, the average activity generated by the stochastic model can be written as  $\mathbf{x}(t) = \mathbb{E}[\mathbf{y}(t)]$ . Given equal initial states  $\mathbf{x}(0) = \mathbf{y}(0)$ ,  $\forall i \in \mathcal{V} : \sum_j a_{ij} \leq 1$ , and  $a_{ij} \geq 0$ , it is straightforward to show that this average network state obeys

$$\mathbf{x}(t) = A\mathbf{x}(t-1) \quad (3.3)$$

(see Methods for a formal proof and S2 Result for numerical validation). Equation 3.3 offers a natural intuition: the average behavior of the stochastic model follows linear dynamics and evolves exponentially as a function of time. Such a relationship allows the application of rich mathematical principles of linear dynamical systems to describe average stochastic dynamics of the model.

### 3.4. Results

#### 3.4.1. Network structure constrains cascade duration

As the first step towards uncovering relations between structure and cascading dynamics, we provide mathematical relationships between network topology and cascade duration. First, the Markov representation above makes explicit the relationship between the network  $A$  and the stimulus propagation and discrimination. The process stated in Equation 3.1 determines a unique map from adjacency matrix  $A$  to transition matrix  $T$ . Given an initial distribution of states, i.e., the stimulus patterns,  $\mathbf{p}(0) = [p_0(0); p_1(0); \dots]$ , the fraction of cascades that terminate, thereby going to the absorbing state (Howard, 1971), by time  $t$  is simply given by the first entry of  $\mathbf{p}(t)$ . Conversely, the probability that a cascade is alive at time  $t$  is given by  $P(\text{alive}, t) = 1 - p_0(t)$ .

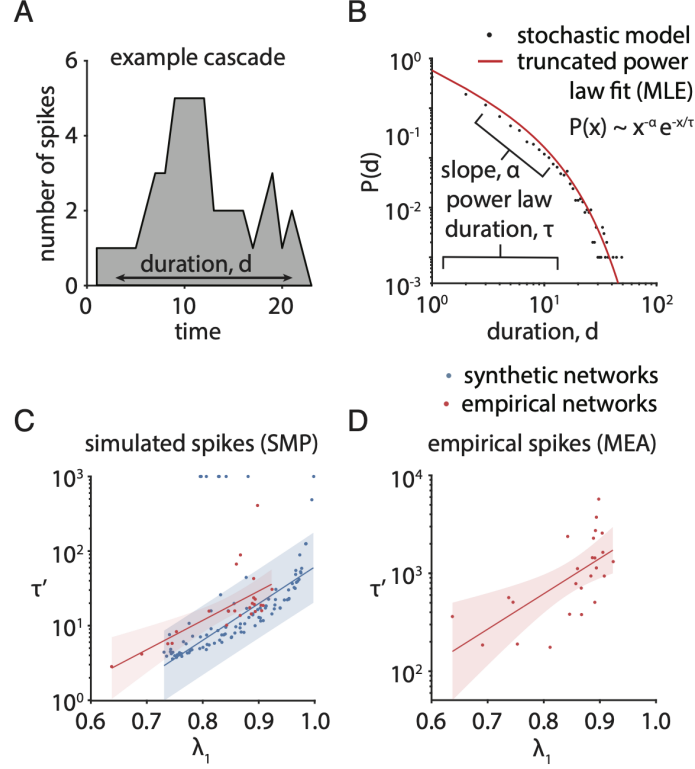
While the Markov representation gives an exact formulation of the stochastic dynamics, the space complexity of the transition matrix is  $O(2^n)$ , making its computation intractable in empirical data with hundreds of neurons. Thus, to more generally describe cascade behavior, we use intuitions grounded in the theory of linear dynamical systems. We can decompose the weight matrix  $A$  into eigenvalues and eigenvectors to identify the elementary modes of activity propagation. Using the dominant eigenvalue  $\lambda_1$  to identify the constraint on the dominant propagation of activity, we can estimate nonlinear, stochastic behavior with a linear system. The dominant eigenvalue  $\lambda_1$  is defined as

$$\lambda_i \in \text{eig}(A) : Av_i = \lambda_i v_i, \quad (3.4)$$

with the maximum absolute value. The dominant eigenvalue  $\lambda_1$  scales the dominant eigenvector  $v_1$ , which constrains the most persistent mode, or vector, of activity propagation (Seung, 1996; Larremore, Shew, and Restrepo, 2011; Larremore, Shew, Ott, et al., 2011), thus quantifying the decay in activity in the network.

To numerically demonstrate the utility of the metric  $\lambda_1$  in explaining cascade duration (Figure 3.2A), we simulated cascades on 104 networks with  $2^8$  nodes for  $10^3$  time steps (see Methods and Supplementary Information for network parameters). Using maximum likelihood estimation (MLE) (Clauset et al., 2009; Alstott et al., 2014), we fit a truncated power law  $p(x) \sim x^{-\alpha} e^{-x/\tau}$  to distributions of cascade duration (Murphy et al., 2019) and computed  $\tau'$  as  $\min(d_{\max}, \tau)$ , bounded by the maximum duration  $d_{\max}$  (Figure 3.2B). (The truncated power law fits the simulated and empirical data better than a power law; see S9 Result.) Intuitively, the metric  $\tau'$  captures the temporal scale in which activity can propagate in a network. We found that  $\tau'$  is monotonically correlated with  $\lambda_1$ , with a Spearman's correlation coefficient  $\rho$  of 0.93 ( $p \approx 0$ ;  $N=104$ ), and that  $\alpha$  has a mean of  $2.0 \pm 0.14$  (standard error; Figure 3.2C) (Bak et al., 1987). Notably, these relations can inform how one would tune the network  $A$  to produce heavy-tailed distributions of cascade duration.

Finally, we empirically tested our predictions in 25 multielectrode array (MEA) recordings of spiking neurons in the mouse somatosensory cortex (Ito et al., 2016) and found similar correlations between network structure and dynamics (Figure 3.2C,D). In the recordings, we binned the spikes into 5ms bins and used MLE to fit a truncated power law and compute  $\tau'$ . To derive  $\lambda_1$ , we calculated an effective connectivity matrix from each recording using first-order vector autoregression (VAR) (Neumaier et al., 2001; Schneider et al., 2001). We found that  $\tau'$  is monotonically correlated with  $\lambda_1$ , as reflected in a Spearman's  $\rho$  of 0.69 ( $p = 1.8 \times 10^{-4}$ ;  $N = 25$ ; Figure 3.2D). Moreover, we can simulate stochastic cascades on the empirically derived networks and find a significant positive correlation between  $\tau'$  and  $\lambda_1$ , with a Spearman's  $\rho$  of 0.68 ( $p = 2.6 \times 10^{-4}$ ;  $N = 25$ ; Figure 3.2C). With a mean



**Figure 3.2: Network topology constrains cascade duration.** **A** Cascade duration is defined as the number of time steps  $t$  between the point at which the first spike occurs after a time step of quiescence, and the point at which the last spike occurs, followed by a time step of quiescence. **B** The distribution of cascade duration can be described by a truncated power law, where parameter  $\alpha$  indicates the log-log slope of the initial distribution and  $\tau$  indicates the duration of the power law on the distribution. **C** In simulations of the stochastic McCulloch-Pitts (SMP) model, the dominant eigenvalue  $\lambda_1$  of synthetic (blue) and empirical (red) networks monotonically scales  $\tau'$  with Spearman's  $\rho$  of 0.93 ( $p = 0$ ;  $N = 104$ ) and 0.68 ( $p = 2.6 \times 10^{-4}$ ;  $N = 25$ ), respectively. Intuitively, the metric  $\tau'$  captures the temporal scale in which activity can propagate in a network. Simulations are run for  $10^3$  time steps. **D** In 25 multielectrode (MEA) recordings, the dominant eigenvalue  $\lambda_1$  of empirical networks monotonically scales  $\tau'$  with Spearman's  $\rho$  of 0.69 ( $p = 1.8 \times 10^{-4}$ ;  $N = 25$ ).

$\alpha$  of  $2.3 \pm 0.1$ , these recordings range in their proximity to criticality (see Supplementary Information for their exponent relations), yet their dynamics are all well-described by their network structures. All together, these results demonstrate the dependence of the temporal scale of activity propagation on the network structure of neural systems.

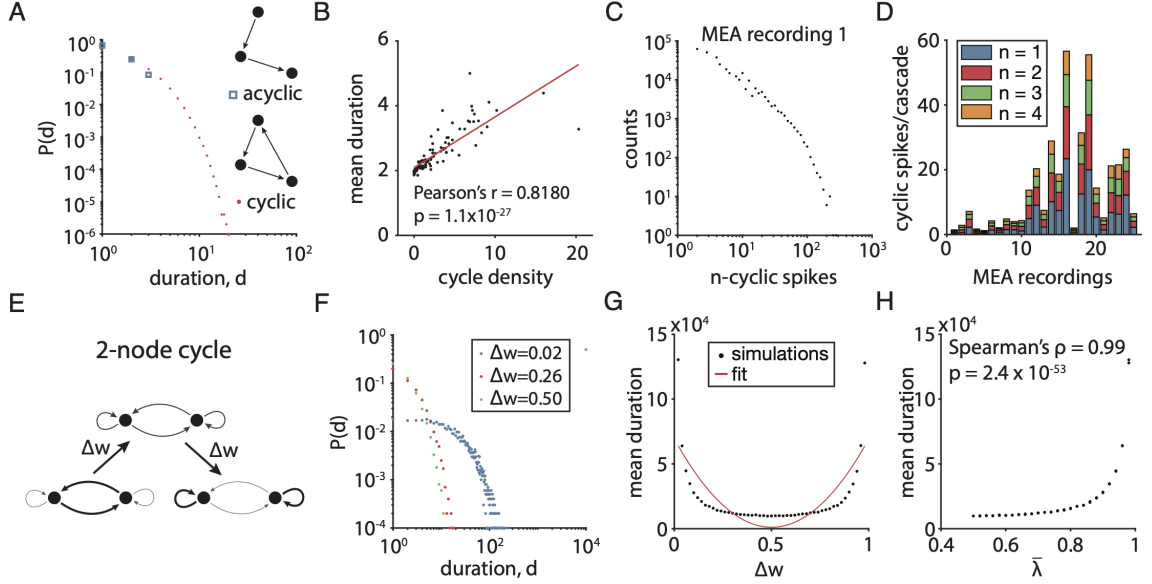


### 3.4.2. Local network structures: cycles

Having demonstrated in the previous section that cascade duration can be predicted from the network structure, we next turn to a deeper examination of which specific features of a network’s topology and geometry can support a heavy-tailed distribution of cascade duration. Note that we use the phrase *network topology* to indicate the arrangement of binary edges and we use the phrase *network geometry* to indicate the distribution of edge weights (Bassett et al., 2013). The two candidate features that we consider are (i) the presence of cycles and (ii) the strength of connections in cycles. We will study these features through a rewiring process on an initial set of edges.

We begin by noting that cycles support temporally extended cascades. Given a single initial stimulus or spontaneous spike, a cascade can have a duration greater than the number of nodes in the graph if and only if there exists at least one cycle in the network. We demonstrate this simple intuition with an acyclic 3-node network and a cyclic 3-node network, where each edge in both networks has a weight of 0.5 (Figure 3.3A). In simulations of  $10^4$  cascades, we found that the acyclic network produces a maximum cascade duration of 3 time steps, as expected. In contrast, using the same number of simulations on the cyclic network, we found the much greater maximum cascade duration of 13 time steps.

Next, we show that the cascade duration scales monotonically with the prevalence of cycles in a network as measured by cycle density, which we define as the number of simple cycles divided by the number of connected edges (Figure 3.3B). To study the effect of cycle density, we begin with a 10-node, directed acyclic graph and randomly rewire each edge with probability  $p$  to a different target node. The directed acyclic graph has the maximum number of edges; that is, the weight matrix is an upper triangular matrix without the diagonal entries. By sweeping over rewiring probabilities  $p = \{0.0, 0.1, 0.2, \dots, 1.0\}$ , we generated networks with different numbers of simple cycles, but the same number of edges and same edge weights of  $\frac{1}{n}$ . For each  $p$ , we simulated  $10^4$  cascades with a maximum duration of



**Figure 3.3: Cycles and strong connections facilitate long cascades.** **A** Distribution of cascade duration in acyclic and cyclic networks (all edges have weight 0.5). In  $10^6$  trials, the maximum cascade durations for the acyclic and cyclic networks are 3 and 11 time steps, respectively. **B** Networks with higher cycle density have longer cascades. We randomly rewired a directed acyclic graph to produce networks of varying cycle density. Cycle density is the number of simple cycles divided by the number of edges. **C** Distribution of  $n$ -cyclic spikes in MEA recording 1 with 5ms bins. An  $n$ -cyclic spike occurs when node  $i$  fires and then fires again after  $n$  time bins. **D** Neurons in mouse somatosensory cortex fire in cycles with small refractory periods. Plot shows distribution of the average number of  $n$ -cycles observed per cascade ( $9.3 \times 10^4 \pm 6.8 \times 10^3$  cascades for 25 recordings; standard error). **E** A schematic of a 2-node network. We redistributed the weights from the 2-node cycle to self-loops by  $\Delta w$ . **F** Distributions of cascade duration for  $\Delta w = 0.02, 0.26$ , and  $0.50$  in the 2-node cycle. **G** Cycles with strong connections, at either  $\Delta w \rightarrow 0$  or  $\Delta w \rightarrow 1$ , extend the mean duration of cascades that do not reach fixed point **1** (quadratic fit:  $y = (2.7 \times 10^5)x^2 - (2.7 \times 10^5)x + (6.9 \times 10^4)$ ). **H** Mean eigenvalue  $\bar{\lambda}$  tracks a network geometry's capacity for long-lasting cascades that do not reach the fixed point **1**.

$10^4$ , and we measured the slope of the linear tail of the distribution on a log-log plot. In these simulations, we found that as a network is rewired to contain more cycles, the average cascade duration increases (Pearson's correlation coefficient  $r = 0.82$ ,  $p = 1.0835 \times 10^{-27}$ ;  $N = 30$ ; Figure 3.3B). These examples illustrate the more general rule that networks containing cycles can support longer cascades and can extend the tail of the distribution of cascade duration.

Importantly, while structural cycles have been experimentally observed (Wang, Markram, et al., 2006; Lefort et al., 2009; Ko et al., 2011), we here empirically validate that activity can actually propagate through cycles. A key potential constraint for cyclical activity propagation is a large refractory period, which can impede such activity even if cycles are structurally present (Michiels van Kessenich et al., 2016). Hence, using the same 25 MEA recordings of spiking cortical neurons as employed previously, we measured the extent of cyclical activity by quantifying the occurrence of  $n$ -cyclic spikes, a phenomenon which occurs in a cascade when a neuron spikes again after  $n$  time bins of its previous spike. We found that on average, 1-, 2-, 3-, and 4-cyclic spikes occur  $14.5 \pm 3.1$  times per cascade (with an average of  $(9.3 \times 10^4) \pm (6.8 \times 10^3)$  cascades for 25 recordings, standard errors; Figure 3.3C,D). With a 5ms bin width, these cyclical activity patterns are within biophysical limits (Connors et al., 1990). Collectively, these results suggest that cyclical activity propagation is not impeded by refractory periods and indeed occurs frequently in living neuronal systems.

### 3.4.3. Local network structures: connection strength

We now turn to a consideration of the distribution of edge weights. To maximize the specificity of our inferences and to generally build our intuition, we constrained ourselves initially to simple networks that only contain a small cycle (a 2-node cycle) or that also contain one relatively larger cycle (a 4-node cycle; see Supplementary Information). We probed the role of weight distributions in the dynamics of the network by placing the strongest weights on edges on one cycle and by placing the weakest weights on edges not on that cycle. Specifically, in both the 2-node and 4-node cycle networks for each simulation, we took the strong weights initially placed on the cycle and redistributed some of their weight by  $\Delta w$  to randomly chosen edges that are not part of the original cycle (i.e.,  $w_{strong,new} := w_{strong,old} - \Delta w$  and  $w_{weak,new} := w_{weak,old} + \Delta w$ ; Figure 3.3E). Upon these new networks, we simulated the stochastic model. We found that as the weight on the original cycle is continuously redistributed away from the initial cycle and throughout the network, we observe fewer and fewer cascades of long duration (Figure 3.3F,G).

Across empirical studies (Beggs and Plenz, 2003; Petermann et al., 2009; Hahn et al., 2010; Friedman et al., 2012; Poil et al., 2012; Lombardi, Herrmann, Plenz, et al., 2014; Bellay et al., 2015; Shew, Clawson, et al., 2015; Ponce-Alvarez et al., 2018), the distributions of avalanche duration have been described by power law functions, where the exponent is known as the lifetime. Typical values vary from -1.0 to -2.6. We seek to show how cycle density and edge weights in cycles together explain the topological and geometric differences in the networks underlying the various distributions of cascade duration. As we redistributed edge weight more uniformly in the networks, we found that mean duration of terminated cascades increases (Figure 3.3F,G). Furthermore, as we redistributed away from the uniform geometry, continuously increasing the range of edge weights, we again observed more and more cascades of long duration. These observations underscore the tight coupling between the range of edge weights, and the heavy-tailed nature of the distribution of cascade duration.

Lastly, we seek to determine whether the distribution of edge weights along cycles contributing to cascade duration is captured by eigenvalue analysis. Towards this goal, we employed the same perturbative numerical experiments on the networks. Specifically, we found that as the weights of the original cycle are redistributed evenly to the alternate cycle (and *vice versa*), the mean duration of cascades increases monotonically with the average of eigenvalues of the network (Spearman’s rank correlation coefficients  $\rho = 0.99$ ,  $p = 2.4 \times 10^{-53}$  and  $r = 0.46$ ,  $p = 9.0 \times 10^{-4}$ , respectively;  $N = 30$ ; Figure 3.3H). Because the dominant eigenvalues of the networks in these simulations are all equal to 1, the average of eigenvalues provide a more descriptive estimation of activity propagation. Thus, this result demonstrates how network geometry can more subtly constrain cascade duration by determining the strength of non-dominant eigenmodes.

#### 3.4.4. Node-specific dynamics

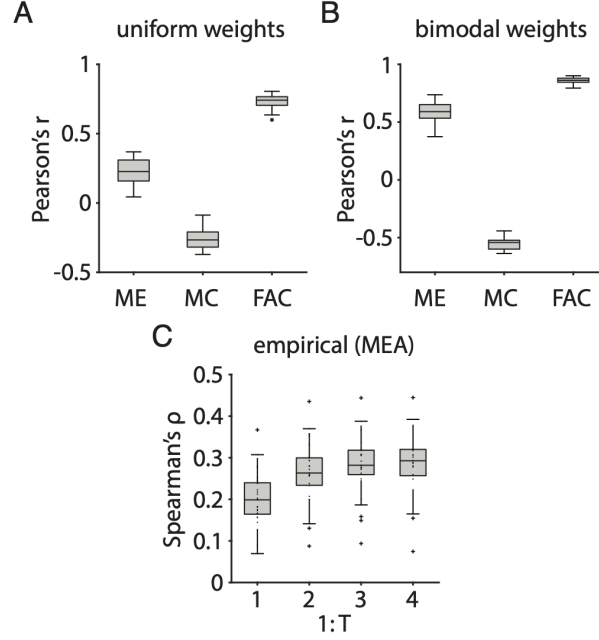
Even within a single network architecture, the range of cascade dynamics can vary depending on the nodes that are stimulated, either spontaneously as the initial state of a cascade or

exogenously through input. Further, cascades follow precise activity patterns that are stable for hours (Beggs, 2004). Thus, we now consider the role of the stimulus pattern on cascade dynamics. We extend our eigenvalue analysis to estimate the role of a stimulus pattern on stochastic cascade dynamics by calculating the magnitude of the eigenprojection of the stimulus pattern. Because the average dynamics are explained by linear systems theory, we then use network control theory to more accurately predict how stimulation of individual nodes alters the dynamics of cascades.

*The eigenprojection of the stimulus pattern.* As a natural extension of the dominant eigenvalue analysis, we first tested whether the magnitude of the eigenprojection of the stimulus pattern could predict cascade dynamics. Given a stimulus  $\mathbf{y}(0)$ , the eigendecomposition of the weight matrix  $A$  into  $A = PDP^{-1}$  yields  $\mathbf{c} = P^{-1}\mathbf{y}(0)$  as the coefficients of the eigenmode excitation of  $\mathbf{y}(0)$ . The components of  $\mathbf{c}$  determine how much the stimulus  $\mathbf{y}(0)$  projects onto the eigenvectors of  $A$  and describes the modes of average activity propagation through the network  $A$ . Then, as a predictor for mean duration, we can compute the 1-norm, or the sum of absolute values, of the eigenprojection of the stimulus pattern scaled by the corresponding eigenvalues  $\boldsymbol{\lambda}$ ,

$$|\mathbf{c} \cdot \boldsymbol{\lambda}|_1. \quad (3.5)$$

We numerically test the eigenprojection metric by simulating cascades on a 100-node, weighted random network. The mean duration of cascades generated from the stimulation of a single node was significantly positively correlated with the magnitude of the eigenprojection (Pearson’s correlation coefficient  $r = 0.34$ ,  $p = 4.5 \times 10^{-4}$ ). To determine the generalizability of these findings, we expanded our simulation set to include 30 random instantiations of networks with the same parameters. In this broader dataset, we found that the Pearson’s correlation coefficient was highly variable (median  $r = 0.23$ ; Figure 3.4A). Thus, we can weakly estimate the role of a stimulus pattern on cascade dynamics with eigenvalue analysis.



**Figure 3.4: Network controllability is tightly linked with cascade duration.** **A** Pearson's correlation coefficients between mean duration of cascades from the stimulation of individual nodes and controllability measures of the respective nodes. Controllability measures include the magnitude of the eigenprojection (ME), modal controllability (MC), and finite average controllability (FAC). The networks here are 30 random instantiations of weighted random graphs, each with 100 nodes and a density of around 0.2. **B** The same plot as in panel a except with a bimodal distribution of weights—with 10% of connections normally distributed with a mean of 0.9 and 90% of connections with a mean of 0.1, all with a standard deviation of 0.1, before weight normalization. **C** Controllability measurements in spiking neurons in mouse somatosensory cortex predict cascade duration. Spearman's correlation between the duration of each cascade and mean finite average controllability of neurons active in its first  $T$  time bins (5ms bins) for 25 MEA recordings. See Supplementary Information for individual plots. The box-plot elements, center, bottom and top edges, whiskers, "+" symbols, indicate respectively, the median, 25th and 75th percentiles, extremes, and outliers. Points are outliers if they are greater than  $q_3 + 1.5 \times (q_3 - q_1)$  or less than  $q_1 - 1.5 \times (q_3 - q_1)$ , where  $q_n$  is the  $n^{\text{th}}$  quartile. Extremes are the most extreme data that are not outliers.

*Network control theory.* To more accurately predict the role of a stimulus pattern on cascade dynamics, we adopt the recently developed metrics of average and modal controllability from network control theory (Pasqualetti et al., 2014). We hypothesized that these metrics, previously applied to large-scale brain networks (Gu, Pasqualetti, et al., 2015; Tang, Giusti, et al., 2017), predicts cascade duration since network control necessitates activity. In the

same set of simulations reported above, we compared the mean cascade duration to the finite average controllability of each node, defined as

$$\text{Trace}(W_K), \quad (3.6)$$

where  $W_K = \sum_{\tau=0}^F A^\tau B_K (A^\tau B_K)^\top$  is the finite controllability Gramian (see Methods). Intuitively, average controllability is the magnitude of the impulse response of the system when stimulating a node. We observed that the mean cascade duration and finite average controllability were significantly positively correlated (Pearson’s correlation coefficient  $r = 0.79$ ,  $p = 2.7 \times 10^{-22}$ ). In contrast, modal controllability was not strongly correlated with mean cascade duration (Pearson’s correlation coefficient  $r = -0.12$ ,  $p = 0.24$ ; see Methods for mathematical definition). Intuitively, modal controllability is a heuristic to determine how well a node produces activity patterns that other nodes cannot easily produce. To determine the generalizability of these findings, we expanded our simulation set to include 30 random instantiations of networks with the same parameters. In this broader dataset, we observed consistent effects (median Pearson’s correlation coefficient  $r = 0.74$  and  $r = -0.27$  for finite average controllability and modal controllability, respectively; Figure 3.4A). In comparing the predictions from linear control theory with the predictions from eigendecomposition, we note that finite average controllability is consistently more strongly correlated with the mean cascade duration than the magnitude of the eigenprojection.

Interestingly, networks with the same topological parameters as above, but with a bimodal distribution of weights show even stronger correlations between network control statistics and cascade dynamics (Figure 3.4B). Such a weight distribution reduces variance in the stochastic process, which intuitively can serve to strengthen the correlation. We observed that the mean cascade duration and finite average controllability were significantly positively correlated (Pearson’s correlation coefficient  $r = 0.87$ ,  $p = 3.2 \times 10^{-32}$ ). Modal controllability became strongly negatively correlated with mean cascade duration (Pearson’s correlation

coefficient  $r = -0.50$ ,  $p = 9.2 \times 10^{-8}$ ). Again to determine the generalizability of these findings, we expanded our simulation set to include 30 random instantiations of networks with the same parameters. We observed consistent effects (mean Pearson’s correlation coefficients between mean cascade duration and finite average controllability, modal controllability, and magnitude of eigenprojection were  $r = 0.86$ ,  $r = -0.54$ , and  $r = 0.59$ , respectively; Figure 3.4B). Again we note that finite average controllability is consistently more strongly correlated with the mean cascade duration than the magnitude of the eigenprojection. These simulations suggest that the skewed weight distributions, as identified in the previous section as network motifs that support long cascades, may strengthen the relationship between network control and network dynamics. Collectively, the results illustrate that the stimulus patterns and the network must be tailored for each other to produce the desired neural dynamics.

Finally, we tested these predictions in empirical data and find that controllability of the initial states is correlated with cascade duration (Figure 3.4C). In each recording from the same MEA data used earlier from spiking neurons in the mouse somatosensory cortex, we calculated the mean finite average controllability of all nodes active in the first  $\{1...T\}$  time bins of each cascade. Mean finite average controllability is monotonically correlated with the duration of each cascade with a median Spearman’s  $\rho = 0.20$  for  $T = 1$  (largest p-value =  $1.9 \times 10^{-33}$ ) and  $\rho = 0.26$  for  $T = 2$  (largest p-value =  $7.9 \times 10^{-49}$ ; see Methods for number of neurons in empirical data). It is important to remember that the cascades are stochastic and cannot be predicted deterministically. Thus, it is notable to find any correlation between mean finite average controllability and cascade duration in empirical data.

### 3.4.5. Cascade duration allows network discriminability and stimulus recovery

If certain network topologies and stimulus patterns can produce long-lasting cascades consistent with avalanche dynamics, what role can lasting cascades contribute to information



processing? Intuitively, one cannot recover information about stimuli from cascades that have already terminated. For lasting cascades, network states can be discriminated and can also provide information about stimuli. Such delayed recovery of stimuli can allow the associative learning of stimuli across temporal delays (Goldman-Rakic, 1995; Durstewitz et al., 2000; Eriksson et al., 2015). The intuition that lasting cascades allow network discriminability can be formalized mathematically via Equation 3.2 and 3.3. Then, with simulations, we test the intuition that cascade dynamics support stimulus recoverability.

In the Markov formulation, the discrimination between network states  $\mathbf{y}(t)$  propagated from stimulus  $\mathbf{y}(0) = \mathbf{s}^i$  and from  $\mathbf{y}(0) = \mathbf{s}^j$  depends upon the similarity between probability vectors  $\mathbf{p}_i(t) = T^t \mathbf{s}^i$  and  $\mathbf{p}_j(t) = T^t \mathbf{s}^j$ . For quickly decaying systems,  $\mathbf{p}_i(t)$  and  $\mathbf{p}_j(t)$  will both have a high probability of being in the zero state  $\mathbf{s}^1$ , inherently reducing discriminability. Hence, the architecture of the network  $A$  constrains the amount of persisting activity that permits discrimination of the initial spiking distribution  $\mathbf{p}(0)$ .

*Network discriminability.* To analytically show the relationship between cascade duration and discriminability, we first define network discriminability as the Euclidean distance between two states  $d(\mathbf{y}_1(t), \mathbf{y}_2(t))$  in  $n$ -dimensional space. Recall that  $\mathbb{E}[\mathbf{y}(t)] = \mathbf{x}(t)$  for stimulus  $\mathbf{x}(0)$  from Equation 3.3. Then, given two stimuli,  $\mathbf{x}_1(0)$  and  $\mathbf{x}_2(0)$ , we can calculate the expected network discriminability as the distance between the expected network states  $d(\mathbf{x}_1(t), \mathbf{x}_2(t))$  at time  $t$ . Given that the dominant eigenvalue  $\lambda_1 < 1$ , then  $\mathbf{x}(t)$  approaches the zero vector  $\mathbf{0}$  as  $t$  approaches  $\infty$ . As described in previous sections, the decay in activity is constrained by the dominant eigenvalue of the network and by the finite average controllability of the individual node being stimulated. Thus, the rate at which both  $\mathbf{x}_1(t)$  and  $\mathbf{x}_2(t)$  decay to  $\mathbf{0}$  determines the rate at which  $d(\mathbf{x}_1(t), \mathbf{x}_2(t))$  approaches  $d(\mathbf{0}, \mathbf{0})$  where discriminability between two network states is zero.

*Stimulus recovery.* To numerically show the relationship between cascade duration and stimulus recoverability, we first define stimulus recoverability as the mutual information  $I(S; Y_t)$  between stimulus patterns  $s \in S$  and network states  $y \in Y_t$  at time  $t$  (see Methods

for details and Figure 3.5A-D for an intuitive schematic). Similar to discriminability, mutual information between the stimuli and network states decreases with shorter cascade duration because the Shannon entropy of the network states decreases. To probe this relation formally, we simulated cascades with 100-node networks from 4 different graph topologies with 30 instantiations of each graph type. Consistent with our intuition, we observe that mutual information is maintained longer when cascades last longer on average (Figure 3.5E). We then quantified the decay in mutual information by first performing linear regression on the mutual information as a function of time for the first 10 time steps. By calculating the Pearson’s correlation coefficient between the slope of linear regression and the mean cascade duration, we found that for all four graph topologies, mutual information decays faster when the propagation of activity also decays faster (Figure 3.5F). Collectively, these results demonstrate that stimulus recoverability is maintained longer when the cascades generated by stimulus patterns last longer.

To link information retention back to network structure, we assessed the relation between stimulus recoverability and the sum of eigenvalues of each network. Using the same 100-node networks from 4 different graph topologies with 30 instantiations of each graph type, we found a significant positive correlation between the average decay rate in mutual information and the sum of eigenvalues, implying that network structure supports the retention of information within the network (Pearson correlation coefficient  $r = 0.92$ ,  $p = 1.8 \times 10^{-49}$ ; Figure 3.5G). Moreover, while all networks had similar parameters, each graph type generated distinct ranges of decay rates and sums of eigenvalues, suggesting that certain graph types may be better suited for information retention than others (Figure 3.5H). In particular, we observe lower decay rates in mutual information and lower sum of eigenvalues in the weighted random and modular graphs, than in the random geometric and Watts-Strogatz graphs. Collectively, these findings demonstrate the interplay among network architecture, network dynamics, and information processing.

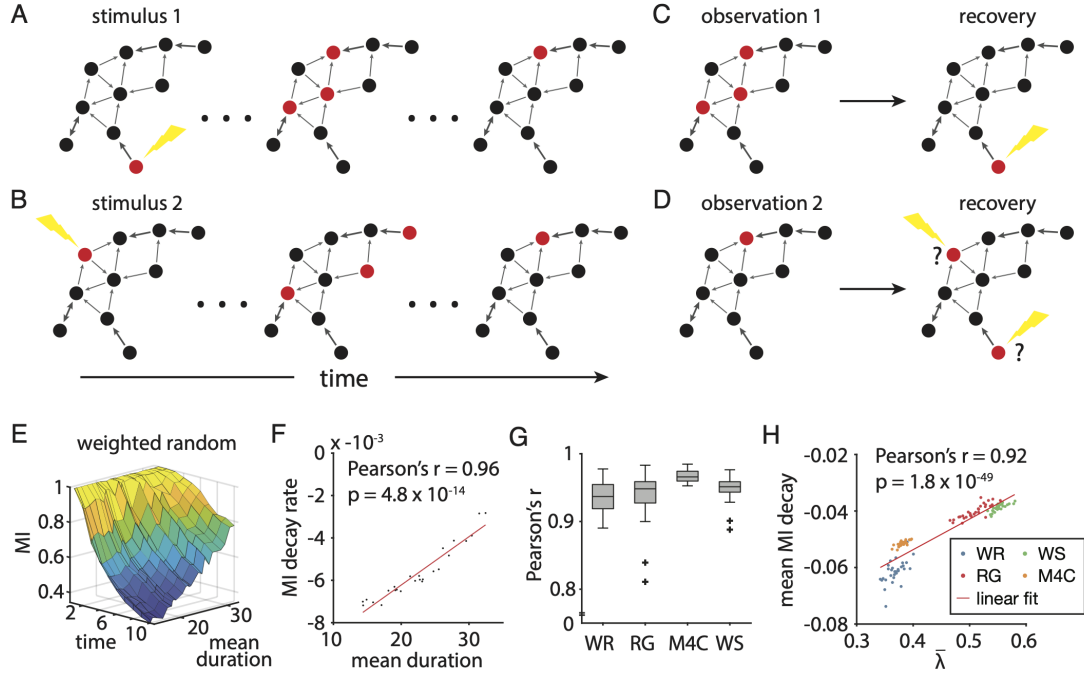


Figure 3.5: **A stimulus can be well-recovered when it generates long-lasting cascades.** **A-B** A schematic showing two cascades triggered by different stimuli. **C** Recovery of the stimulus using an observation of a network state during a cascade. **D** Failed recovery of the stimulus. **E** Decay in mutual information (MI) over time. When activity from a stimulus pattern lasts longer, mutual information also persists for longer for a weighted random graph. **F** The linear decay rate of mutual information over the first 10 time steps plotted against the mean cascade duration in the example weighted random graph from panel e. **G** The Pearson correlation coefficients between the linear slope of decay in mutual information over time and the mean cascade duration for four graph types: a weighted random graph (WR), a random geometric graph (RG), a modular graph with 4 communities (M4C), and a Watts-Strogatz graph (WS). The boxplot shows data from 30 instantiations of each graph type, each network containing 100 nodes and characterized by a fractional connectivity of around 0.05. The whiskers extend to the extreme data points not considered outliers, and the outliers are plotted individually using the “+” symbol. **H** The mean decay rate in mutual information for a network is correlated with the sum of eigenvalues of the network (Pearson’s correlation coefficient  $r = 0.92$ ,  $p = 1.8 \times 10^{-49}$ ;  $N = 120$ ). For all networks, we used a fractional connectivity of 0.05 to show a wide range of decay rates in mutual information (see Supplementary Information for simulations with other fractional connectivities).

### 3.5. Discussion

Neural systems display strikingly rich dynamics that harbor the marks of a complex underlying network architecture among units, from the small scale of individual neurons to the

large scale of columns and areas (Wang, Chen, et al., 2013; Nigam et al., 2016). Cascades are a quintessential example of such dynamics, and, when they are close to a critical regime, are thought to allow for a diverse range of computations (Beggs and Plenz, 2003; Haldeman et al., 2005; Kinouchi et al., 2006; Shew, Yang, Petermann, et al., 2009). Yet, precisely how a neuronal network’s structure supports stochastic dynamics and the computations that can arise therefrom remains unclear. Here, we seek to provide clarity using both precise analysis of mathematical formulations and statistically rigorous assessments of numerical experiments. We consider a generalized stochastic spiking model and demonstrate that the time-averaged activity of this model can be treated as a linear dynamical system. From this observation, we derive intuitions for how network structure, which estimates the patterns in synaptic interactions, constrains cascade duration. In subsequent numerical experiments and empirical validation, we use eigendecomposition and network control theory appropriate for linear dynamical systems to describe how network structure and the stimulus pattern together determine the manner in which a stimulus propagates through the network during a neural cascade. We identify strongly connected cycles, which have been widely empirically observed, as prevalent network motifs that promote long cascade duration in neuronal networks. Finally, we use mutual information to demonstrate that long-lasting cascades can serve as a mechanism to allow for temporally delayed recovery of desired patterns of stimulation. Broadly, our work blends dynamical systems theory, network control theory, information theory, and computational neuroscience to address the wide gap in the field’s current understanding of the relations between architecture, dynamics, and computation.

### **3.5.1. Biophysical implications of results**

The biophysical implications of the results demonstrated here are threefold. The first implication is on the scale of a local neuronal population of hundreds of neurons. On this scale, we showed that the dominant eigenvalue of a network scales the distribution in the duration of cascades. From a broad perspective, this result shows that complex behavior of a neuronal population can be described by the collective pairwise interactions between

neurons. The second implication is on the scale of a handful of neurons. On this scale, neurons form cycles through which spikes can propagate. This result demonstrates that the extensive empirical observation of bidirectionally connected neurons (Wang, Markram, et al., 2006; Lefort et al., 2009; Ko et al., 2011) is integral to propagating activity in a network. Importantly, a refractory period, which can limit cyclic activity if large enough (Michiels van Kessenich et al., 2016), does not prohibit a cyclic propagation of activity, at least at a temporal scale of 5ms and longer. The third implication is on the scale of individual neurons. On this scale, the results of the eigenprojection and controllability analyses show that a neuron propagates activity for longer if it has a large magnitude of the eigenprojection or if it has high controllability. In a neuronal population, different neurons can have different roles in performing computations depending on the topology (Faber et al., 2019). Our results suggest that high controllability neurons may serve as “broadcasting neurons” which, upon activation, propagate activity for a long duration to the entire network.

### 3.5.2. Linear form of stochastic network dynamics

Because of the inherently stochastic nature of neuronal cascades, many previous studies have simply inferred properties about the underlying network through statistical methods (Beggs and Plenz, 2003; Lombardi, Herrmann, Perrone-Capano, et al., 2012). An important innovation in this study was the demonstration that the time-averaged activity of the stochastic system has an equivalent form as a linear dynamical system. In real neuronal systems, dynamics are non-linear, which most likely accounts for the difference in range of  $\tau'$  in Figures 3.2C and 3.2D. Such linear estimation of the dynamics makes available powerful computational tools in matrix and linear systems theory, and allowed us to capitalize on recent advances in network control (Liu et al., 2011; Pasqualetti et al., 2014). Network control theory is a formal approach to modeling, predicting, and tuning the response of a networked system to exogenous input, and has been recently applied to neural systems at both the cellular (Yan et al., 2017; Wiles et al., 2017; Towlson et al., 2018) and regional (Gu, Pasqualetti, et al., 2015; Tang, Giusti, et al., 2017; Cornblath et al., 2018; Jeganathan

et al., 2018) scales (for a recent review, see (Tang and Bassett, 2018)). In these previous efforts, linear dynamics have been assumed, whereas here such dynamics have been proven, to be relevant for the neural system under study. Extensions of linear systems analysis, such as observability (Chen, 1998) and optimal control (Taylor et al., 2015; Betzel et al., 2016; Gu, Betzel, et al., 2017), follow immediately from this work and could provide added insights into other dynamical and computational properties of neural networks. Finally, it would be of interest to directly probe the effects of stimulation patterns defined by network controllability statistics on information transmission *in vitro* or behaviors *in vivo*, following work in a similar vein in large-scale human neuroimaging (Medaglia et al., 2018; Muldoon et al., 2016; Stiso et al., 2019; Khambhati et al., 2018).

### 3.5.3. Topological constraints on dynamics and computation

Proving formally that network topology affects dynamics and computation is important, but can be further complemented by providing intuitions regarding the specific features of a network topology that are most relevant, thus explaining and guiding experimental results. The identification of functionally relevant features of networked systems has a long history in molecular biology (Alon, 2007), with notable efforts identifying structural motifs in transcription regulation networks (Shen-Orr et al., 2002), protein-protein interaction networks (Yeager-Lotem et al., 2004), and cellular circuits (Hart et al., 2012), which are thought to arise spontaneously under evolutionary pressures (Kashtan et al., 2005). Significantly extending prior statistical efforts in large-scale connectomes (Sporns et al., 2004), here we demonstrate that specific structural motifs in the form of strongly connected cycles are topological features that support long cascade dynamics. These structural motifs form elementary units or building blocks of the network that can be combined to create connectivity architectures that produce certain dynamical behaviors (Shimono et al., 2014; Nigam et al., 2016). Other theoretical studies have also found strongly and bi-directionally connected neurons as motifs that produce long-lasting memory (Brunel, 2016), potentially as a mechanism for attractor dynamics (Hopfield, 1982). Importantly, empirical studies have shown that the network mo-

tifs identified here are observed in both cortical microcircuits (Wang, Markram, et al., 2006; Lefort et al., 2009; Ko et al., 2011; Markram, 1997; Song, Sjöström, et al., 2005; Perin et al., 2011) and macrocircuits (Sizemore, Giusti, et al., 2018). Future work is needed to better understand the rules by which neurons connect to one another, and to determine whether those rules serve to increase the memory capacity of cortical networks. It would also be interesting in the future to determine whether higher-order structural motifs, such as those accessible to tools from algebraic topology (Giusti et al., 2016; Sizemore, Phillips-Cremins, et al., 2018), might also play a role in the relationships between topology, dynamics, and computation (Sizemore, Giusti, et al., 2018; Reimann et al., 2017).

#### **3.5.4. Information theory as a performance measure**

To measure information retention, we use mutual information between stimulus patterns and network states, but it only captures a certain aspect of information processing. Mutual information, originally developed to study communication channels (Shannon, 1948), has proven to be a powerful tool for the study of information transmission in avalanching neural networks (Beggs and Plenz, 2003; Shew, Yang, Yu, et al., 2011). While previous studies of neuronal avalanches use power law statistics that suggest criticality as the theoretical link between dynamics and information processing (Beggs and Plenz, 2003; Bertschinger et al., 2004; Haldeman et al., 2005; Kinouchi et al., 2006; Shew, Yang, Petermann, et al., 2009; Shriki et al., 2013; Shew, Clawson, et al., 2015), we take a more mechanistic approach embedded in dynamical systems theory to study the relationships between network structure, dynamics, and mutual information. While there is substantial evidence that cortical networks frequently operate near a critical point (Friedman et al., 2012; Fontenele et al., 2019), this is not always the case (Priesemann et al., 2014; Touboul et al., 2010; Touboul et al., 2017); we therefore did not assume that all activity took the form of critical avalanches. Our more generic approach allowed us to develop a framework that would apply all cascades, critical or not. Despite its utility in studying information channels, mutual information is unlikely to be the only useful performance measure for a neural system, given

the numerous purported computations of cortical networks (Shew, Yang, Petermann, et al., 2009; Timme et al., 2016). Indeed, the explanation posited here for the prevalence of strongly connected neurons does not account for the information faculties of the rest of the neural system. Such considerations compel further investigation into how network structure supports other types of information processing accessible to other information theoretic measures.

### **3.5.5. Methodological considerations**

A few remarks are warranted on the topic of linear dynamics in neural systems. Linear dynamics accurately predicts stochastic, cascade dynamics, and its rich mathematical properties have been used to study neural dynamics in many organisms across a wide range of temporal and spatial scales (Liu et al., 2011; Gu, Pasqualetti, et al., 2015; Kim et al., 2018; Yan et al., 2017). At the neuronal level, however, neural dynamics are non-linear (Hodgkin et al., 1952). Efforts analytically demonstrating properties about non-linear systems are more limited (Motter, 2015), and thus, further study is required to more thoroughly demonstrate the relationships shown here in a non-linear system.

### **3.5.6. Future directions**

In closing, we note that the natural direction in which to take this work will be to consider other types of information processing and to identify network structures and neuronal dynamics of different cell types that produce complex network dynamics which in turn support such computations. Here, we demonstrate that the rich mathematical properties of linear systems can reveal insights into the complex dynamics of non-linear, non-deterministic neural systems. In the future, we can further apply this theory to cascading and other neural systems to ask questions about networks, their dynamics, and their computations. It would be apt to apply this framework to cortical networks from functional, structural, and effective connectivities and measure memory performance in terms of the network topology and dynamics. It would be interesting to measure differences in memory performance across brain



regions, and to test for relationships between topological features and performance. Third and finally, studying well-known network learning rules—such as Hebbian plasticity (Hebb, 1949) and spike-timing dependent plasticity (Song, Miller, et al., 2000)—in a dynamical systems and information theoretic framework may shed further light on the functional purpose of these rules.

### 3.6. Methods

#### 3.6.1. Synthetic network generation

We use five different commonly studied graph models from network science in our analyses (Wu-Yan et al., 2018). The first graph model is the *Weighted Random Graph* model (WRG), which is a weighted version of the canonical Erdős-Rényi model. The weight of an edge is distributed as a geometric distribution with probability of success  $p$ . Second, we use a *Random Geometric* model (RG) that is embedded in a unit cube, where the edge weights are equal to the inverse of the Euclidean distance between two nodes. We kept only a fraction of the shortest edges in order to achieve a desired edge density  $p$ . Third, we use a *Modular Graph with 4 Communities* model (MD4). Pairs of nodes within communities have an edge density of 0.8, and nodes across communities are connected to achieve a desired edge density of  $p$ . The edges of nodes in the same community and across communities are weighted according to a geometric distribution with probability of success  $p$  and  $1 - p$ , respectively. Fourth, we use a *Watts-Strogatz* model (WS). The model builds a ring lattice and then uniformly rewires the network, creating a small-world architecture with a random probability of  $r = 0.1$ . Fifth, we use a *Hierarchical Modular Graph* (HM). The model generates a directed network with  $m$  hierarchical levels of modules with size  $s$ , and connection density decays as  $1/E^n$ . See Supplementary Information for a summary of the graph models used in simulations.

### 3.6.2. Empirical network generation

For analysis of an empirical system, we use publicly available data derived from spiking neurons in the mouse somatosensory cortex (Ito et al., 2016). The data contain 25 recordings, most of which possess hundreds of neurons (min: 98, max: 594, mean: 309, total: 7735). Each recording is 60 minutes long and was acquired at a sampling rate of 20 kHz. The recordings were acquired from organotypic slice cultures by multielectrode arrays (MEAs), each with 512 electrodes on a 1mm-by-2mm area.

We obtain empirical networks by calculating the effective connectivity of spiking neurons in the empirical data (Ito et al., 2016). On the spike trains of each recording, we first bin the spike trains into 5ms bins. With 5ms bins, we capture almost all action potential propagation and synaptic transmission in the array area (Friedman et al., 2012; Shimono et al., 2014; Nigam et al., 2016). In a previous study (see Supplemental Figure 7 in (Shimono et al., 2014)), the authors use the identical multielectrode array data and show that the distribution of delays, measured with transfer entropy, falls largely within 5ms. Then, using the ARfit software package for MATLAB (Schneider et al., 2001), we perform vector autoregression (VAR) for the autoregressive (AR) model:

$$\mathbf{y}(t) = w + \sum_{l=1}^p A_l \mathbf{y}(t-l),$$

where  $\mathbf{y}(t)$  is a vector representing the number of spikes for each neuron at time  $t$ ,  $w$  is a vector of intercept terms, and the matrices  $A_1, \dots, A_p \in \mathbb{R}^{m \times m}$  are the coefficient matrices of the AR model (Neumaier et al., 2001). With 5ms time bins, each term of the VAR model captures synaptic delays within 5ms. For all empirical networks, negative edge weights are allowed to capture inhibition (Hayashi et al., 2018). We set the lower and upper bounds for the model order,  $p_{min}$  and  $p_{max}$ , to 1 and 4, respectively. After selecting an optimal model

order  $p_{opt}$  using Schwarz’s Bayesian Criterion (Schwarz, 1978), we compute the effective connectivity  $A$  as the sum of the coefficient matrices  $A_1, \dots, A_p \in^{n \times n}$  of the VAR model over the model orders such that for the elements  $a_{ij}$  in  $A$  and  $a_{l,ij}$  in  $A_l$ ,  $a_{ij} = \sum_{l=1}^{p_{opt}} a_{l,ij}$ . The effective connectivity  $A$  is equivalent to a linear system, which in turn equals the stochastic McCulloch-Pitts neurons averaged across trials given the constraints laid out in Equation 3.3. One advantage of using an autoregressive model to build an effective connectivity network, compared to, for example, a transfer entropy network, is that one can directly use linear systems theory to analyze the linearized dynamics of the network.

### 3.6.3. Network analysis

We use three sets of weight distributions: a uniform distribution, a truncated normal distribution, and a bimodal distribution. In some simulations, however, we explicitly set the weights to particular values. In a uniform distribution of weights, we set all weights equal to 1 and normalize each row. In a truncated normal distribution, we set the non-zero weights to the upper half of a truncated normal distribution. A truncated normal distribution of weights has been widely observed both in a theoretical context with synaptic plasticity and in the experimental literature (Brunel et al., 2004; Iyer et al., 2013; Pehlevan et al., 2017). Lastly, we use a skewed, bimodal distribution with a few connections centered at a normal distribution with a large mean and most other connections centered at a normal distribution with a small mean. Bimodal distributions occur theoretically in the context of additive synaptic plasticity (Rossum et al., 2000), and positively skewed distributions have been observed experimentally (Markram et al., 1997; Feldmeyer et al., 1999; Chen et al., 2010). Our skewed, bimodal distributions combine these two observations by having a few strong connections. All weights are static and do not change with time  $t$ . See the Supplementary Information for network parameters.

To calculate the cycle density of a graph, we compute the number of simple cycles divided by the number of connected edges. A simple cycle is defined as the set of edges in a closed

walk with no repetitions of vertices and edges, other than the starting and ending vertex. The number of simple cycles was calculated using the `networkx` software package (version 2.1) on Python (version 3.7.3).

#### 3.6.4. Simulating the Stochastic McCulloch-Pitts model

We model cascades as spikes propagating through a recurrent network (see Mathematical Framework). For computational tractability, we set a maximum time step  $K$  for the simulations. The simulated spike counts  $\mathbf{y}(t)$  are stored as a  $n$ -by- $K$  matrix. All simulations and calculations were run on MATLAB (version 2018a) provided by The MathWorks, Inc.

#### 3.6.5. Stimulus pattern generation

We investigate the propagation of activity through a network initiated by stimulus patterns. The stimulus pattern is set as the initial state  $\mathbf{y}(0)$  or  $\mathbf{x}(0)$  of a network and then propagated forward in time according to either stochastic or linear dynamics, respectively. In our study, we consider two ways to generate stimulus patterns. In the analysis of cascade duration and controllability, we stimulate individual nodes by creating a set of vectors in which the  $i^{th}$  element of the  $i^{th}$  vector is set at 1 and all other elements are set at 0. In the mutual information analysis, we create a set of column vectors such that their finite average controllability values evenly span the range of controllability values (see later section of this Methods for definition of finite average controllability). In each of the  $\frac{n}{m} = 25$  vectors, we choose  $m = 4$  nodes from  $n = 100$  total nodes to stimulate such that each node that we select is increasing in its finite average controllability value. Because finite average controllability is highly correlated with cascade duration, such input vectors will evenly span the possible duration of cascades.

### 3.6.6. Characterizing distributions of cascade duration

We characterized the distributions of cascade duration using a truncated power law. We used maximum likelihood estimation to estimate the power law with exponential cutoff  $P(x) \sim x^{-\alpha} e^{x/\tau}$  (Clauset et al., 2009; Alstott et al., 2014). The exponent  $\tau$  describes the value of  $x$  at which the exponential cuts off the tail of the power law duration. To avoid overgeneralizing the extent of the power law, we bound  $\tau$  by the maximum duration of  $x$  and indicate this bound value as  $\tau' = \min(\tau, \max(x))$ .

### 3.6.7. Mutual information calculation to probe stimulus recovery

To measure the capacity of a network to transfer information during a cascade, we calculated the mutual information  $I(X; Y)$ , which quantifies the amount of information, in bits, that one random variable  $X$  reveals about another random variable  $Y$ . Here, the two random variables of interest are the initial network state  $\mathbf{y}(0)$ , where  $\mathbf{y}(0)$  is a stimulus  $\mathbf{s}^i \in S$ , and the network states  $\mathbf{y}(t) \in Y_t$  at a later time  $t$ . With mutual information, we measure the amount of information that the network states  $Y_t$  at time  $t$  reveal about the stimulus patterns  $S$ . For each stimulus pattern  $\mathbf{s}^i$ , we simulated 1,000 cascades where  $P(\mathbf{s}^i) = P(\{\mathbf{s}^j | j \neq i\}) = 0.5$ . (See earlier Methods section on “Stimulus pattern generation”.) All mutual information calculations were run using the MIToolbox (v3.0.1) for MATLAB (<https://github.com/Craigacp/MIToolbox>).

In the analysis of the relationship between the average cascade duration and the mutual information, we quantify the decay in mutual information over time. We also calculate the correlation between the decay rate of the mutual information and the predicted mean cascade duration. For this latter calculation, first we perform a linear regression of the decay in mutual information with respect to time. Then, we calculate the Pearson correlation coefficient between the slope of the linear regression and the mean cascade duration.

### 3.6.8. Estimation by linear dynamical systems

We prove by induction that linear dynamics estimates average behavior of the stochastic model, i.e.,  $\mathbb{E}[y_j(t)] = x_j(t)$ , given the same initial conditions  $\mathbf{y}(0) = \mathbf{x}(0)$ . At  $t = 0$ , both  $y_j(0)$  and  $x_j(0)$  are set as the stimulus pattern, and so,  $\mathbb{E}[y_j(0)] = x_j(0)$ . Now, assume  $\mathbb{E}[y_j(t-1)] = x_j(t-1)$ , and see that  $x_j(t) = a_j^T \mathbf{x}(t-1) = a_j^T \mathbb{E}[\mathbf{y}(t-1)] = \mathbb{E}[a_j^T \mathbf{y}(t-1)] = \mathbb{E}[y_j(t)]$  and thus,  $\mathbb{E}[y_j(t)] = x_j(t)$ . To demonstrate this relation numerically, we take the average cascades that begin with the same initial state by taking the mean of  $y_i^k(t)$  for all cascades  $k$  at each time step  $t$ . All cascades start with the same initial condition  $\mathbf{y}(0)$ . We perform numerical validation in the Supplement Information.

### 3.6.9. Eigenvalue analysis

In our analysis of networks, we decompose the weight matrix  $A$  into eigenvalues and eigenvectors. Such an eigendecomposition is formalized as

$$A = PDP^{-1}, \quad (3.7)$$

where  $P$  is a matrix of eigenvectors as columns and  $D$  is a diagonal matrix of corresponding eigenvalues. We calculate the absolute value of the eigenvalue with the largest absolute value as the dominant eigenvalue  $\lambda_1$ .

When the row sum  $\sum_j a_{ij}$  is greater than 1, the linear dynamical system does not equal the expected value of the stochastic model. However, the eigenvalue analyses can still be useful in describing average stochastic behavior. In particular, when  $\lambda_1 > 1$ , the state  $\mathbf{x}(t)$  of the linear dynamical system can explode exponentially. While the state  $\mathbf{y}(t)$  of a stochastic model with the same parameters does not similarly explode exponentially, it is bound by 1 for each neuron and reaches a fixed point at  $\mathbf{1}$ . In this case, the states of both models,  $\mathbf{x}(t)$  and  $\mathbf{y}(t)$ , cannot reach quiescence at  $\mathbf{0}$  and thus have infinite cascade duration.

### 3.6.10. Network control theory and controllability statistics

Network control theory is a formulation of control theory for networks of interacting components. This formulation typically consists of a set of  $n$  component nodes  $\mathcal{V} = \{1, \dots, n\}$ , where the vector  $\mathbf{x}(t) \in^n$  represents the state of node activities at time  $t \geq 0$ . These nodes are connected by a set of edges  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ , where the adjacency matrix  $A \in^{n \times n}$  has elements  $a_{ij}$  as the strength of the connection from node  $j$  to node  $i$ . Here, *control* typically refers to a set of  $k$  inputs  $\mathbf{u}(t) \in^k$  at time  $t \geq 0$  that drive the evolution of system states according to  $B \in^{n \times k}$ . In linear control theory, the system states evolve as

$$\mathbf{x}(t+1) = A\mathbf{x}(t) + B\mathbf{u}(t). \quad (3.8)$$

### 3.6.11. Finite average controllability

Motivated by a desire to understand how network architecture affects its control properties, recent work iterates network-based metrics for control of such linear systems (Pasqualetti et al., 2014). Particularly germane to our discussion of cascade duration is *average controllability* (Kailath, 1980; Gu, Pasqualetti, et al., 2015), defined as the  $H_2$  norm of the system's infinite average controllability given by

$$\text{Trace}[W_K] = \text{Trace} \left[ \sum_{\tau=0}^{\infty} A^\tau B B^T A^{T\tau} \right]. \quad (3.9)$$

Here, we set  $B$  as a binary column vector where vector elements corresponding to the nodes of interest are set to 1 and the remaining vector elements are set to 0; this formulation represents an impulse of magnitude 1 to the nodes of interest. The finite average controllability (FAC) is similarly defined by taking the sum to some finite positive integer  $F$  instead of infinity, and represents the norm of the system's impulse response over  $F$  time steps. Because cascades are expected to last for a finite number of time steps, we use  $F = 100$  in the main text, and in the supplement we show that larger and smaller values of  $F$  produce

similar results.

### 3.6.12. Modal controllability

Another network-based control metric we use here is *modal controllability* (Pasqualetti et al., 2014; Gu, Pasqualetti, et al., 2015). While modal controllability was originally formulated for symmetric matrices, here we extend the definition to include asymmetric matrices. To do this, we take the absolute value of both the eigenvalues and the eigenvector components, which can be complex numbers in an asymmetric matrix. Thus, we define the version of modal controllability of node  $i$  for asymmetric matrices as

$$\phi_i = \sum_{j=1}^n (1 - |\lambda_j|^2) |v_{ij}|^2. \quad (3.10)$$

### 3.6.13. Finite average controllability of initial states

To predict the duration of a cascade, we can calculate the finite average controllability of an initial state  $\mathbf{y}(0)$  defined as the finite average controllability averaged over the nodes that are active in the initial state,

$$\text{FAC}(\mathbf{y}(0)) = \frac{1}{|\mathbf{y}(0)|} \sum_{i \in \{i | y_i(0)=1\}} \text{FAC}_i. \quad (3.11)$$

In the same way, we also calculate the finite average controllability in empirical cascades in the first  $\{1 \dots T\}$  time bins, averaging over the active neurons in those bins.

## 3.7. Code and data availability

All code for simulations and analysis is publicly available at <https://github.com/harangju/cascades>. All data that support the findings of this study are available in the Open Science Framework with the identifier doi:10.17605/OSF.IO/TW69H.



## REFERENCES

- Adolphs, Ralph (Mar. 1, 2003). “Cognitive neuroscience of human social behaviour.” In: *Nature Reviews Neuroscience* 4, 165 EP.
- Alon, U (2007). “Network motifs: theory and experimental approaches.” In: *Nat Rev Genet* 8.6, pp. 450–461.
- Alstott, Jeff, Ed Bullmore, and Dietmar Plenz (Jan. 2014). “powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions.” In: *PLOS ONE* 9.1, pp. 1–11. DOI: 10.1371/journal.pone.0085777.
- Bak, Per, Chao Tang, and Kurt Wiesenfeld (July 1987). “Self-organized criticality: An explanation of the  $1/f$  noise.” In: *Physical Review Letters* 59.4, pp. 381–384.
- Bassett, Danielle S et al. (Mar. 2013). “Robust detection of dynamic community structure in networks.” In: *Chaos* 23.1, p. 013142. DOI: 10.1063/1.4790830.
- Beggs, J. M. (2004). “Neuronal Avalanches Are Diverse and Precise Activity Patterns That Are Stable for Many Hours in Cortical Slice Cultures.” In: *Journal of Neuroscience* 24.22, pp. 5216–5229. DOI: 10.1523/JNEUROSCI.0540-04.2004.
- Beggs, John M. and Dietmar Plenz (2003). “Neuronal Avalanches in Neocortical Circuits.” In: *Journal of Neuroscience* 23.35, pp. 11167–11177. DOI: 10.1523/JNEUROSCI.23-35-11167.2003.
- Bellay, Timothy et al. (July 2015). “Irregular spiking of pyramidal neurons organizes as scale-invariant neuronal avalanches in the awake state.” In: *eLife* 4. Ed. by Frances K Skinner, e07224. DOI: 10.7554/eLife.07224.
- Ben-Yishai, R, R L Bar-Or, and H Sompolinsky (1995). “Theory of orientation tuning in visual cortex.” In: *Proceedings of the National Academy of Sciences* 92.9, pp. 3844–3848. DOI: 10.1073/pnas.92.9.3844.
- Bertschinger, Nils and Thomas Natschläger (2004). “Real-Time Computation at the Edge of Chaos in Recurrent Neural Networks.” In: *Neural Computation* 16.7, pp. 1413–1436. DOI: 10.1162/089976604323057443.

- Betzel, R F et al. (2016). “Optimally controlling the human connectome: the role of network topology.” In: *Sci Rep* 6, p. 30770.
- Brunel, Nicolas (Apr. 11, 2016). “Is cortical connectivity optimized for storing information?” In: *Nature Neuroscience* 19, 749 EP.
- Brunel, Nicolas et al. (Apr. 2004). “Optimal Information Storage and the Distribution of Synaptic Weights: Perceptron versus Purkinje Cell.” In: *Neuron* 43.5, pp. 745–757. DOI: 10.1016/j.neuron.2004.08.023.
- Chen, Chi-Tsong (1998). *Linear System Theory and Design*. 3rd. New York, NY, USA: Oxford University Press, Inc. ISBN: 0-19-511777-8.
- Chen, Wei et al. (Jan. 2010). “A few strong connections: optimizing information retention in neuronal avalanches.” In: *BMC neuroscience* 11, pp. 3–3. DOI: 10.1186/1471-2202-11-3.
- Clauset, A., C. Shalizi, and M. Newman (2009). “Power-Law Distributions in Empirical Data.” In: *SIAM Review* 51.4, pp. 661–703. DOI: 10.1137/070710111.
- Connors, Barry W. and Michael J. Gutnick (1990). “Intrinsic firing patterns of diverse neocortical neurons.” In: *Trends in Neurosciences* 13.3, pp. 99–104. DOI: 10.1016/0166-2236(90)90185-D.
- Cornblath, E J et al. (2018). “Sex differences in network controllability as a predictor of executive function in youth.” In: *Neuroimage* 188, pp. 122–134.
- Daie, Kayvon, Mark S. Goldman, and Emre R. F. Aksay (Mar. 2015). “Spatial Patterns of Persistent Neural Activity Vary with the Behavioral Context of Short-Term Memory.” In: *Neuron* 85.4, pp. 847–860. DOI: 10.1016/j.neuron.2015.01.006.
- Durstewitz, Daniel, Jeremy K. Seamans, and Terrence J. Sejnowski (Nov. 1, 2000). “Neurocomputational models of working memory.” In: *Nature Neuroscience* 3, 1184 EP.
- Eriksson, Johan et al. (2015). “Neurocognitive Architecture of Working Memory.” In: *Neuron* 88.1, pp. 33–46. DOI: 10.1016/j.neuron.2015.09.020.
- Faber, Samantha P. et al. (2019). “Computation is concentrated in rich clubs of local cortical networks.” In: *Network Neuroscience* 3.2, pp. 384–404. DOI: 10.1162/netn\_a\_00069.

- Feldmeyer, Dirk et al. (Apr. 1999). “Reliable synaptic connections between pairs of excitatory layer 4 neurones within a single ‘barrel’ of developing rat somatosensory cortex.” In: *The Journal of Physiology* 521.1, pp. 169–190. DOI: 10.1111/j.1469-7793.1999.00169.x.
- Fiete, Ila R. et al. (Mar. 2010). “Spike-Time-Dependent Plasticity and Heterosynaptic Competition Organize Networks to Produce Long Scale-Free Sequences of Neural Activity.” In: *Neuron* 65.4, pp. 563–576. DOI: 10.1016/j.neuron.2010.02.003.
- Fontenele, Antonio J. et al. (May 2019). “Criticality between Cortical States.” In: *Phys. Rev. Lett.* 122.20, p. 208101. DOI: 10.1103/PhysRevLett.122.208101.
- Friedman, Nir et al. (May 2012). “Universal Critical Dynamics in High Resolution Neuronal Avalanche Data.” In: *Phys. Rev. Lett.* 108.20, p. 208102. DOI: 10.1103/PhysRevLett.108.208102.
- Gireesh, Elakkat D. and Dietmar Plenz (2008). “Neuronal avalanches organize as nested theta- and beta/gamma-oscillations during development of cortical layer 2/3.” In: *Proceedings of the National Academy of Sciences* 105.21, pp. 7576–7581. DOI: 10.1073/pnas.0800537105.
- Giusti, C, R Ghrist, and D S Bassett (2016). “Two’s company, three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data.” In: *J Comput Neurosci* 41.1, pp. 1–14.
- Goldman-Rakic, P. S. (1995). “Cellular basis of working memory.” In: *Neuron* 14.3, pp. 477–485. DOI: 10.1016/0896-6273(95)90304-6.
- Gu, S, R F Betzel, et al. (2017). “Optimal trajectories of brain state transitions.” In: *Neuroimage* 148, pp. 305–317.
- Gu, Shi, Fabio Pasqualetti, et al. (2015). “Controllability of structural brain networks.” In: *Nature Communications* 6.1, p. 8414. DOI: 10.1038/ncomms9414.
- Hahn, Gerald et al. (2010). “Neuronal Avalanches in Spontaneous Activity In Vivo.” In: *Journal of Neurophysiology* 104.6, pp. 3312–3322. DOI: 10.1152/jn.00953.2009.

- Haldeman, Clayton and John M. Beggs (Feb. 2005). “Critical Branching Captures Activity in Living Neural Networks and Maximizes the Number of Metastable States.” In: *Phys. Rev. Lett.* 94.5, p. 058101. DOI: 10.1103/PhysRevLett.94.058101.
- Hart, Y et al. (2012). “Design principles of cell circuits with paradoxical components.” In: *Proc Natl Acad Sci U S A* 109.21, pp. 8346–8351.
- Hayashi, Ayako, Takashi Yoshida, and Kenichi Ohki (Dec. 2018). “Cell Type Specific Representation of Vibro-tactile Stimuli in the Mouse Primary Somatosensory Cortex.” In: *Frontiers in neural circuits* 12, pp. 109–109. DOI: 10.3389/fncir.2018.00109.
- Hebb, Donald (1949). *The Organization of Behavior: a Neuropsychological Theory*. Oxford, England: Wiley.
- Hodgkin, A L and A F Huxley (Aug. 1952). “A quantitative description of membrane current and its application to conduction and excitation in nerve.” In: *The Journal of physiology* 117.4, pp. 500–544.
- Honey, Christopher J. et al. (2007). “Network structure of cerebral cortex shapes functional connectivity on multiple time scales.” In: *Proceedings of the National Academy of Sciences* 104.24, pp. 10240–10245. DOI: 10.1073/pnas.0701519104.
- Hopfield, J J (1982). “Neural networks and physical systems with emergent collective computational abilities.” In: *Proceedings of the National Academy of Sciences* 79.8, pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554.
- Howard, R. A. (1971). *Markov Models*. New York, NY, USA: Wiley.
- Ito, Shinya et al. (2016). *Spontaneous spiking activity of hundreds of neurons in mouse somatosensory cortex slice cultures recorded using a dense 512 electrode array*. DOI: 10.6080/K07D2S2F.
- Iyer, Ramakrishnan et al. (Oct. 2013). “The influence of synaptic weight distribution on neuronal population dynamics.” In: *PLoS computational biology* 9.10, e1003248, e1003248–e1003248. DOI: 10.1371/journal.pcbi.1003248.

- Jeganathan, J et al. (2018). “Fronto-limbic dysconnectivity leads to impaired brain network controllability in young people with bipolar disorder and those at high genetic risk.” In: *Neuroimage Clin* 19, pp. 71–81.
- Kailath, Thomas (1980). *Linear systems*. Prentice-Hall information and system science series. Englewood Cliffs, N.J: Prentice-Hall. ISBN: 978-0-13-536961-6.
- Kashtan, N and U Alon (2005). “Spontaneous evolution of modularity and network motifs.” In: *Proc Natl Acad Sci U S A* 102.39, pp. 13773–13778.
- Khambhati, A N et al. (2018). “Predictive control of electrophysiological network architecture using direct, single-node neurostimulation in humans.” In: *Network Neuroscience* [https://www.mitpressjournals.org/doi/abs/10.1162/netn\\_a\\_00089](https://www.mitpressjournals.org/doi/abs/10.1162/netn_a_00089).
- Kim, Jason Z. et al. (2018). “Role of graph architecture in controlling dynamical networks with applications to neural systems.” In: *Nature Physics* 14.1, pp. 91–98. DOI: 10.1038/nphys4268.
- Kinouchi, Osame and Mauro Copelli (Apr. 23, 2006). “Optimal dynamical range of excitable networks at criticality.” In: *Nature Physics* 2, 348 EP.
- Ko, Ho et al. (Apr. 10, 2011). “Functional specificity of local synaptic connections in neocortical networks.” In: *Nature* 473, 87 EP.
- Larremore, Daniel B., Woodrow L. Shew, and Juan G. Restrepo (Jan. 2011). “Predicting Criticality and Dynamic Range in Complex Networks: Effects of Topology.” In: *Phys. Rev. Lett.* 106.5, p. 058101. DOI: 10.1103/PhysRevLett.106.058101.
- Larremore, Daniel B, Woodrow L Shew, Edward Ott, et al. (June 2011). “Effects of network topology, transmission delays, and refractoriness on the response of coupled excitable systems to a stochastic stimulus.” In: *Chaos (Woodbury, N.Y.)* 21.2, pp. 025117–025117. DOI: 10.1063/1.3600760.
- Lefort, Sandrine et al. (2009). “The Excitatory Neuronal Network of the C2 Barrel Column in Mouse Primary Somatosensory Cortex.” In: *Neuron* 61.2, pp. 301–316. DOI: 10.1016/j.neuron.2008.12.020.

- Liu, Yang-Yu, Jean-Jacques Slotine, and Albert-László Barabási (2011). “Controllability of complex networks.” In: *Nature* 473.7346, pp. 167–173. DOI: 10.1038/nature10011.
- Lombardi, F., H. J. Herrmann, C. Perrone-Capano, et al. (May 2012). “Balance between Excitation and Inhibition Controls the Temporal Organization of Neuronal Avalanches.” In: *Phys. Rev. Lett.* 108.22, p. 228703. DOI: 10.1103/PhysRevLett.108.228703.
- Lombardi, Fabrizio, Hans J. Herrmann, Dietmar Plenz, et al. (2014). “On the temporal organization of neuronal avalanches.” In: *Frontiers in Systems Neuroscience* 8, p. 204. DOI: 10.3389/fnsys.2014.00204.
- Markram, H (1997). “A network of tufted layer 5 pyramidal neurons.” In: *Cerebral Cortex* 7.6, pp. 523–533. DOI: 10.1093/cercor/7.6.523.
- Markram, H et al. (Apr. 1997). “Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex.” In: *The Journal of physiology* 500 ( Pt 2) (Pt 2), pp. 409–440.
- McCulloch, Warren S. and Walter Pitts (1990). “A logical calculus of the ideas immanent in nervous activity. 1943.” In: *Bulletin of mathematical biology* 52 1-2, 99–115, discussion 73–97.
- Medaglia, J D et al. (2018). “Network Controllability in the Inferior Frontal Gyrus Relates to Controlled Language Variability and Susceptibility to TMS.” In: *J Neurosci* 38.28, pp. 6399–6410.
- Michiels van Kessenich, L., L. de Arcangelis, and H. J. Herrmann (Aug. 18, 2016). “Synaptic plasticity and neuronal refractory time cause scaling behaviour of neuronal avalanches.” In: *Scientific Reports* 6, 32071 EP.
- Motter, Adilson E (Sept. 2015). “Networkcontrolology.” In: *Chaos (Woodbury, N.Y.)* 25.9, pp. 097621, 097621–097621. DOI: 10.1063/1.4931570.
- Muldoon, S F et al. (2016). “Stimulation-Based Control of Dynamic Brain Networks.” In: *PLoS Comput Biol* 12.9, e1005076.

- Murphy, Kieran A., Karin A. Dahmen, and Heinrich M. Jaeger (Jan. 2019). “Transforming Mesoscale Granular Plasticity Through Particle Shape.” In: *Phys. Rev. X* 9.1, p. 011014. DOI: 10.1103/PhysRevX.9.011014.
- Neumaier, Arnold and Tapio Schneider (2001). “Estimation of parameters and eigenmodes of multivariate autoregressive models.” In: *ACM Transactions on Mathematical Software* 27.1, pp. 27–57. DOI: 10.1145/382043.382304.
- Nigam, Sunny et al. (2016). “Rich-Club Organization in Effective Connectivity among Cortical Neurons.” In: *Journal of Neuroscience* 36.3, pp. 670–684. DOI: 10.1523/JNEUROSCI.2177-15.2016.
- Pasqualetti, F, S Zampieri, and F Bullo (2014). “Controllability Metrics, Limitations and Algorithms for Complex Networks.” In: *IEEE Transactions on Control of Network Systems* 1.1, pp. 40–52.
- Pehlevan, C. and A. Sengupta (2017). “Resource-efficient perceptron has sparse synaptic weight distribution.” In: *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4. DOI: 10.1109/SIU.2017.7960683.
- Perin, Rodrigo, Thomas K. Berger, and Henry Markram (2011). “A synaptic organizing principle for cortical neuronal groups.” In: *Proceedings of the National Academy of Sciences* 108.13, pp. 5419–5424. DOI: 10.1073/pnas.1016051108.
- Petermann, Thomas et al. (2009). “Spontaneous cortical activity in awake monkeys composed of neuronal avalanches.” In: *Proceedings of the National Academy of Sciences* 106.37, pp. 15921–15926. DOI: 10.1073/pnas.0904089106.
- Poil, Simon-Shlomo et al. (2012). “Critical-State Dynamics of Avalanches and Oscillations Jointly Emerge from Balanced Excitation/Inhibition in Neuronal Networks.” In: *Journal of Neuroscience* 32.29, pp. 9817–9823. DOI: 10.1523/JNEUROSCI.5990-11.2012.
- Ponce-Alvarez, Adrián et al. (Nov. 2018). “Whole-Brain Neuronal Activity Displays Crackling Noise Dynamics.” In: *Neuron*. DOI: 10.1016/j.neuron.2018.10.045.

- Priesemann, Viola et al. (2014). “Spike avalanches in vivo suggest a driven, slightly subcritical brain state.” In: *Frontiers in Systems Neuroscience* 8, p. 108. DOI: 10.3389/fnsys.2014.00108.
- Reimann, M W et al. (2017). “Cliques of Neurons Bound into Cavities Provide a Missing Link between Structure and Function.” In: *Front Comput Neurosci* 11, p. 48.
- Rodriguez, Paul and William B Levy (2001). “A model of hippocampal activity in trace conditioning: Where’s the trace?” In: *Behavioral Neuroscience* 115.6, pp. 1224–1238. DOI: 10.1037/0735-7044.115.6.1224.
- Rossum, M. C. W. van, G. Q. Bi, and G. G. Turrigiano (Dec. 2000). “Stable Hebbian Learning from Spike Timing-Dependent Plasticity.” In: *The Journal of Neuroscience* 20.23, p. 8812. DOI: 10.1523/JNEUROSCI.20-23-08812.2000.
- Schneider, Tapio and Arnold Neumaier (Mar. 2001). “Algorithm 808: ARfit—A Matlab Package for the Estimation of Parameters and Eigenmodes of Multivariate Autoregressive Models.” In: *ACM Trans. Math. Softw.* 27.1, pp. 58–65. DOI: 10.1145/382043.382316.
- Schwarz, Gideon (1978). “Estimating the Dimension of a Model.” In: *The Annals of Statistics* 6.2, pp. 461–464.
- Seung, H. S. (Nov. 1996). “How the brain keeps the eyes still.” In: *Proceedings of the National Academy of Sciences* 93.23, p. 13339. DOI: 10.1073/pnas.93.23.13339.
- Shannon, C. E. (1948). “A Mathematical Theory of Communication.” In: *Bell System Technical Journal* 27.3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.
- Shen-Orr, S S et al. (2002). “Network motifs in the transcriptional regulation network of *Escherichia coli*.” In: *Nat Genet* 31.1, pp. 64–68.
- Shew, Woodrow L., Wesley P. Clawson, et al. (2015). “Adaptation to sensory input tunes visual cortex to criticality.” In: *Nature Physics* 11.8, pp. 659–663. DOI: 10.1038/nphys3370.
- Shew, Woodrow L., Hongdian Yang, Thomas Petermann, et al. (2009). “Neuronal Avalanches Imply Maximum Dynamic Range in Cortical Networks at Criticality.” In: *Journal of Neuroscience* 29.49, pp. 15595–15600. DOI: 10.1523/JNEUROSCI.3864-09.2009.



- Shew, Woodrow L., Hongdian Yang, Shan Yu, et al. (2011). “Information Capacity and Transmission Are Maximized in Balanced Cortical Networks with Neuronal Avalanches.” In: *Journal of Neuroscience* 31.1, pp. 55–63. DOI: 10.1523/JNEUROSCI.4637-10.2011.
- Shimono, Masanori and John M. Beggs (Oct. 2014). “Functional Clusters, Hubs, and Communities in the Cortical Microconnectome.” In: *Cerebral Cortex* 25.10, pp. 3743–3757. DOI: 10.1093/cercor/bhu252.
- Shriki, Oren et al. (2013). “Neuronal Avalanches in the Resting MEG of the Human Brain.” In: *Journal of Neuroscience* 33.16, pp. 7079–7090. DOI: 10.1523/JNEUROSCI.4286-12.2013.
- Sizemore, A E, C Giusti, et al. (2018). “Cliques and cavities in the human connectome.” In: *J Comput Neurosci* 44.1, pp. 115–145.
- Sizemore, A E, J E Phillips-Cremins, et al. (2018). “The importance of the whole: Topological data analysis for the network neuroscientist.” In: *Network Neuroscience* Epub Ahead of Print.
- Song, Sen, Kenneth D. Miller, and L. F. Abbott (2000). “Competitive Hebbian learning through spike-timing-dependent synaptic plasticity.” In: *Nature Neuroscience* 3.9, pp. 919–926. DOI: 10.1038/78829.
- Song, Sen, Per Jesper Sjöström, et al. (Mar. 2005). “Highly Nonrandom Features of Synaptic Connectivity in Local Cortical Circuits.” In: *PLOS Biology* 3.3. DOI: 10.1371/journal.pbio.0030068.
- Sporns, O and R Kotter (2004). “Motifs in brain networks.” In: *PLoS Biol* 2.11, e369.
- Stiso, Jennifer et al. (2019). “White Matter Network Architecture Guides Direct Electrical Stimulation through Optimal State Transitions.” In: *Cell Reports* 28.10, 2554–2566.e7. DOI: 10.1016/j.celrep.2019.08.008.
- Tang, E and D S Bassett (2018). “Control of dynamics in brain networks.” In: *Rev. Mod. Phys.* 90, p. 031003.
- Tang, E, C Giusti, et al. (2017). “Developmental increases in white matter network controllability support a growing diversity of brain dynamics.” In: *Nat Commun* 8.1, p. 1252.

- Taylor, P N et al. (2015). “Optimal control based seizure abatement using patient derived connectivity.” In: *Front Neurosci* 9, p. 202.
- Timme, Nicholas M. et al. (May 2016). “High-Degree Neurons Feed Cortical Computations.” In: *PLOS Computational Biology* 12.5, pp. 1–31. DOI: 10.1371/journal.pcbi.1004858.
- Touboul, Jonathan and Alain Destexhe (Feb. 2010). “Can Power-Law Scaling and Neuronal Avalanches Arise from Stochastic Dynamics?” In: *PLOS ONE* 5.2, pp. 1–14. DOI: 10.1371/journal.pone.0008982.
- (Jan. 2017). “Power-law statistics and universal scaling in the absence of criticality.” In: *Phys. Rev. E* 95.1, p. 012413. DOI: 10.1103/PhysRevE.95.012413.
- Towlson, E K et al. (2018). “Caenorhabditis elegans and the network control framework-FAQs.” In: *Philos Trans R Soc Lond B Biol Sci* 373, p. 1758.
- Wang, Xiao-Jing (Sept. 2002). “Probabilistic Decision Making by Slow Reverberation in Cortical Circuits.” In: *Neuron* 36.5, pp. 955–968. DOI: 10.1016/S0896-6273(02)01092-9.
- Wang, Yun, Henry Markram, et al. (Mar. 19, 2006). “Heterogeneity in the pyramidal network of the medial prefrontal cortex.” In: *Nature Neuroscience* 9, 534 EP.
- Wang, Zheng, Li Min Chen, et al. (2013). “The Relationship of Anatomical and Functional Connectivity to Resting-State Connectivity in Primate Somatosensory Cortex.” In: *Neuron* 78.6, pp. 1116–1126. DOI: 10.1016/j.neuron.2013.04.023.
- Watts, Duncan J. and Steven H. Strogatz (June 4, 1998). “Collective dynamics of ‘small-world’ networks.” In: *Nature* 393, 440 EP.
- Wiles, L et al. (2017). “Autaptic Connections Shift Network Excitability and Bursting.” In: *Sci Rep* 7, p. 44006.
- Wu-Yan, Elena et al. (Mar. 9, 2018). “Benchmarking Measures of Network Controllability on Canonical Graph Models.” In: *Journal of Nonlinear Science*. DOI: 10.1007/s00332-018-9448-z.
- Yan, Gang et al. (2017). “Network control principles predict neuron function in the Caenorhabditis elegans connectome.” In: *Nature* 550.7677, pp. 519–523. DOI: 10.1038/nature24056.

Yeger-Lotem, E et al. (2004). “Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction.” In: *Proc Natl Acad Sci U S A* 101.16, pp. 5934–5939.

# Supplementary Information

## 3.8. Supplementary Results

### 3.8.1. Numerical validation of Markov formulation

To numerically assess the constraint on cascade duration, we compared simulations of the network in Figure 3.1A-B to the prediction  $P(\text{alive}, t)$  given by the Markov representation. We observed little difference between the stochastic and predicted dynamics. For each of  $10^6$  trials, we stimulated single neurons at  $t = 1$ , and at each time step (from a maximum of 100), we calculated the fraction of cascades alive and  $P(\text{alive}, t)$ . We found that the root-mean-square error (RMSE) between the Markov chain prediction and the stochastic model was  $1.2 \times 10^{-4}$  (Figure 3.6). To determine the generalizability of our observations, we extended this analysis to an ensemble of 120 networks, separated into 30 instantiations of four different graph topologies chosen for their relevance to neuronal architectures: a weighted random graph, a ring lattice graph, a modular graph with 4 communities, and a Watts-Strogatz graph. For the four graph topologies, we observed that the average RMSEs were less than  $8.5 \times 10^{-3}$ . Taken together, these results indicate a tight link between network structure  $A$  and cascade duration derived from the network dynamics  $T$ .

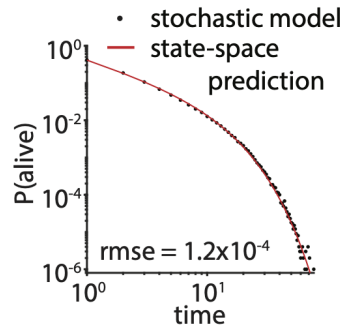


Figure 3.6: **Numerical validation of Markov formulation.** A Markov chain of network states can accurately predict the fraction of active cascades at time  $t$ . In  $10^4$  trials of stimulating neuron 8 in the network in panel b, the root-mean-square error between the state-space prediction and the stochastic model is  $1.2 \times 10^{-4}$ .

### 3.8.2. Numerical validation of the relationship between the linear and stochastic models

To illustrate the relation between the stochastic and the linear model, we perform numerical simulations of the model, and we compare simulated cascades to the average network state estimated by a linear dynamical system. In both cases, we instantiate the dynamics on a weighted random network comprised of 10 nodes (Figure 3.1A-B) (Garlaschelli, 2009). The relevant network parameters for all simulations are listed in Section 3.9.1. We simulate the stochastic model dynamics 1,000 times over 15 time steps starting with the same initial condition  $\mathbf{y}(0)$  (Figure 3.7A). Note that we can also consider this initial condition to be the stimulus. We average the activity at each node and time step  $y_j(t)$  across simulations to generate a numerical estimate of the time-evolution of the average network state (Figure 1B). Then, using the linear dynamical system starting with the same initial condition  $\mathbf{x}(0) = \mathbf{y}(0)$ , we calculated the number of spikes per neuron per time step as an estimate of the average network state (Figure 3.7C). We find that the difference between the states of the linear system and of the stochastic cascading model approaches 0 as a function of trials  $k$  (Figure 3.1D). This convergence is consistent across a range of network sizes for fixed density (Figure 3.1E). These results illustrate the accuracy of the linear estimation of the dynamics of the stochastic model.

### 3.8.3. Random redistribution from 4-node cycles to different sets of edges

In this subsection, we reproduce the results from Figure 3.3I-L in the main text, but with different sets of edges (see Figure 3.8). The results provided here are qualitatively similar to those shown in Figure 3.3.

### 3.8.4. Finite average controllability for different time periods

We demonstrate that the finite average controllability is correlated with the mean cascade duration even with higher values of  $F$ , a parameter reflecting the time period over which

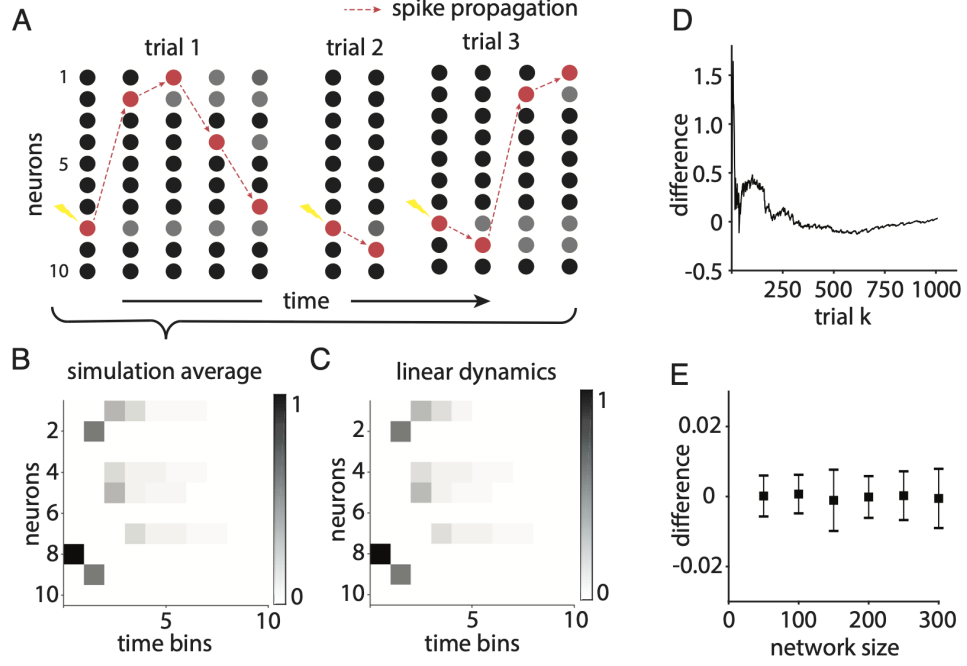


Figure 3.7: **Linear dynamical model predicts average firing rate** **A** Examples of simulations of cascades generated by stimulating neuron 8 in the network in panel b. **B** The activity of each node and time step, averaged over  $10^4$  cascades of stimulating neuron 8 in the network in panel a. **C** Linear dynamics estimates the average spike counts of stochastic simulations in panel d. **D** The difference between linear dynamics and simulation average converges to a steady-state around zero. **E** The differences between average simulated spiking and estimated linear dynamics for weighted random networks of size 50, 100, 150, 200, 250, and 300 nodes, all with fractional connectivity of 0.2. The error bars indicate standard deviations, and the means are  $2.1 \times 10^{-4}$ ,  $-6.8 \times 10^{-5}$ ,  $-9.7 \times 10^{-6}$ ,  $-8.0 \times 10^{-5}$ ,  $5.8 \times 10^{-5}$ , and  $-2.1 \times 10^{-5}$ , respectively.

the system's average controllability is measured. In the literature (Gu et al., 2015), average controllability is defined as  $\text{Trace}(W_K)$  where  $W_K = \sum_{\tau=0}^{\infty} A^{\tau} B_K B_K^T A^{\tau}$ . Because cascades are expected to last for a finite number of time steps, we define finite average controllability as the trace of a finite version of the controllability Gramian,  $W_K = \sum_{\tau=0}^F A^{\tau} B_K B_K^T A^{\tau}$ , as discussed more fully in the Methods subsection of the main text. Intuitively, finite average controllability is the finite impulse response of the system. Here, we show that higher values of  $F$  only increase the correlation between mean cascade duration from a stimulus and the finite average controllability of the stimulus (see Figure 3.9). For readers curious about pragmatic concerns, we note that the estimation of finite average controllability becomes

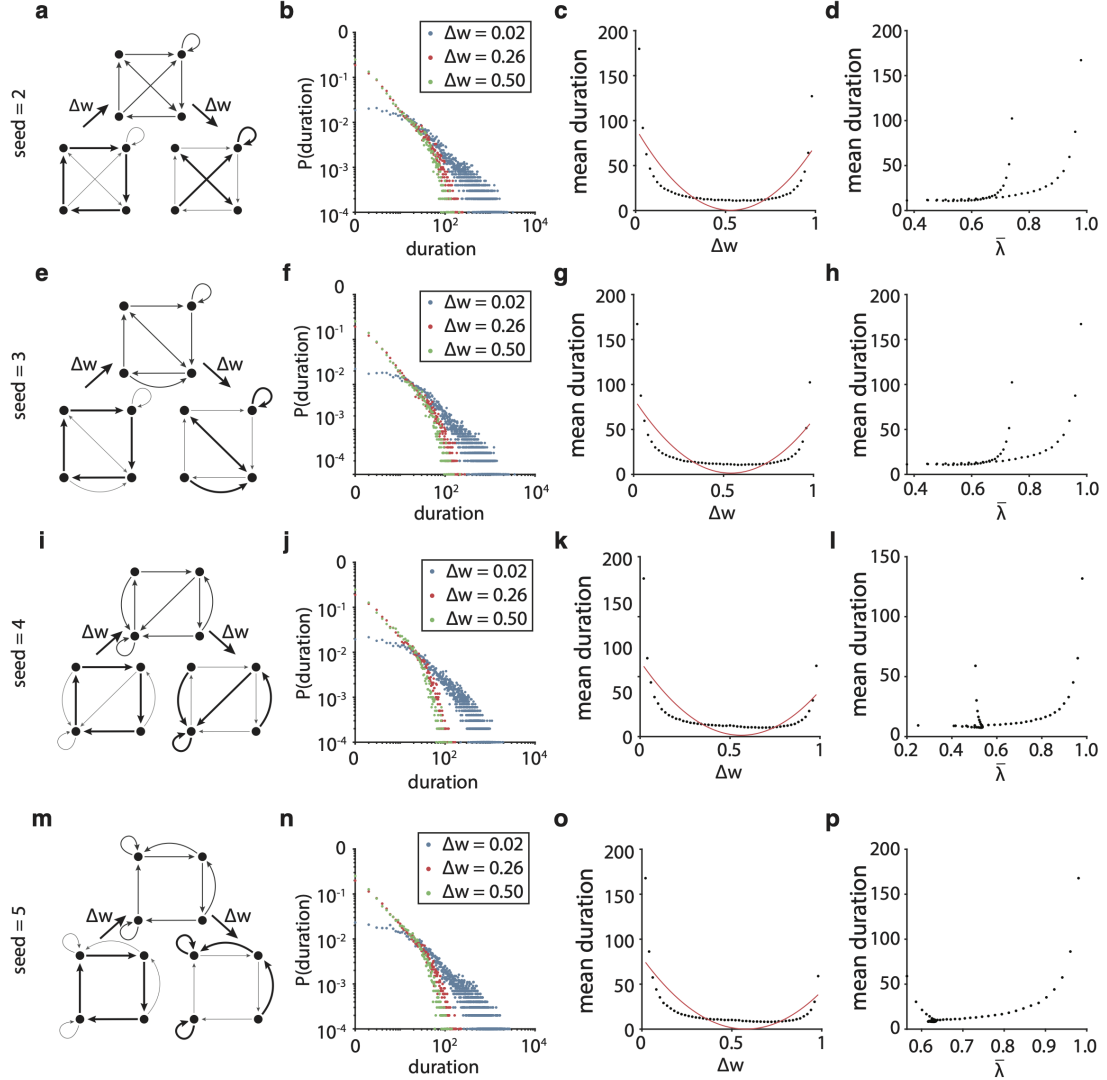


Figure 3.8: **Strong edge weights in cycles produce long cascades.** **A-D** Reproduction of the results from Figure 3.3H-K with a different set of edges generated by a random seed of 2. The subpanels A, B, C, and D here correspond to subpanels H, I, J, and K of Figure 3.3. **E-H** Reproduction of the results from Figure 3.3H-K with a different set of edges generated by a random seed of 3. **I-L** Reproduction of the results from Figure 3.3H-K with a different set of edges generated by a random seed of 4. Panels M-P reproduce the results from Figure 3.3H-K with a different set of edges generated by a random seed of 5.

much more computationally intensive as  $F$  increase.

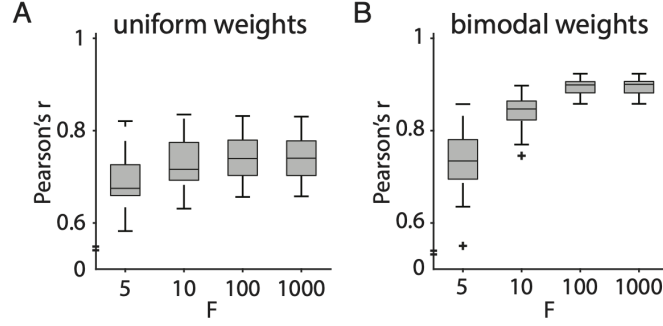


Figure 3.9: **Finite average controllability, estimated for a wide range of  $F$  values, is correlated with mean cascade duration.** **A** The Pearson's correlation coefficient,  $r$ , between the mean duration of cascades from a single-node stimulation and the finite average controllability, for values of  $F$  ranging from 5 to 1,000. The networks here have a weighted random network topology with fractional connectivity of 0.2 and a uniform distribution of weights. **B** The Pearson's correlation coefficient,  $r$ , for the same measurements as in panel A except for a bimodal distribution of weights, as explained in the main text.

### 3.8.5. Mutual information decay calculations for networks with different fractional connectivities

Here we reproduce the results from Figure 3.4E-J in the main text with networks of higher fractional connectivities of 0.1 and 0.2. The networks with higher fractional connectivities shown here display similarly high correlations between (i) the decay in mutual information between stimuli and network state over time and (ii) the mean duration of cascades (see Figure 3.10).

### 3.8.6. Collisions in cascades

One assumption of branching processes is that collisions can be ignored. Some branching process models indeed formulate avalanches as tree structures (Lee et al., 2004). In contrast, both our stochastic model and linear systems allow collisions. To address this discrepancy, we assess the degree to which collisions occur in empirical data, using the same 25 multielectrode array (MEA) recordings described in the Methods subsection. While collisions cannot be measured directly, we measured correlation between the number of neurons active in a time



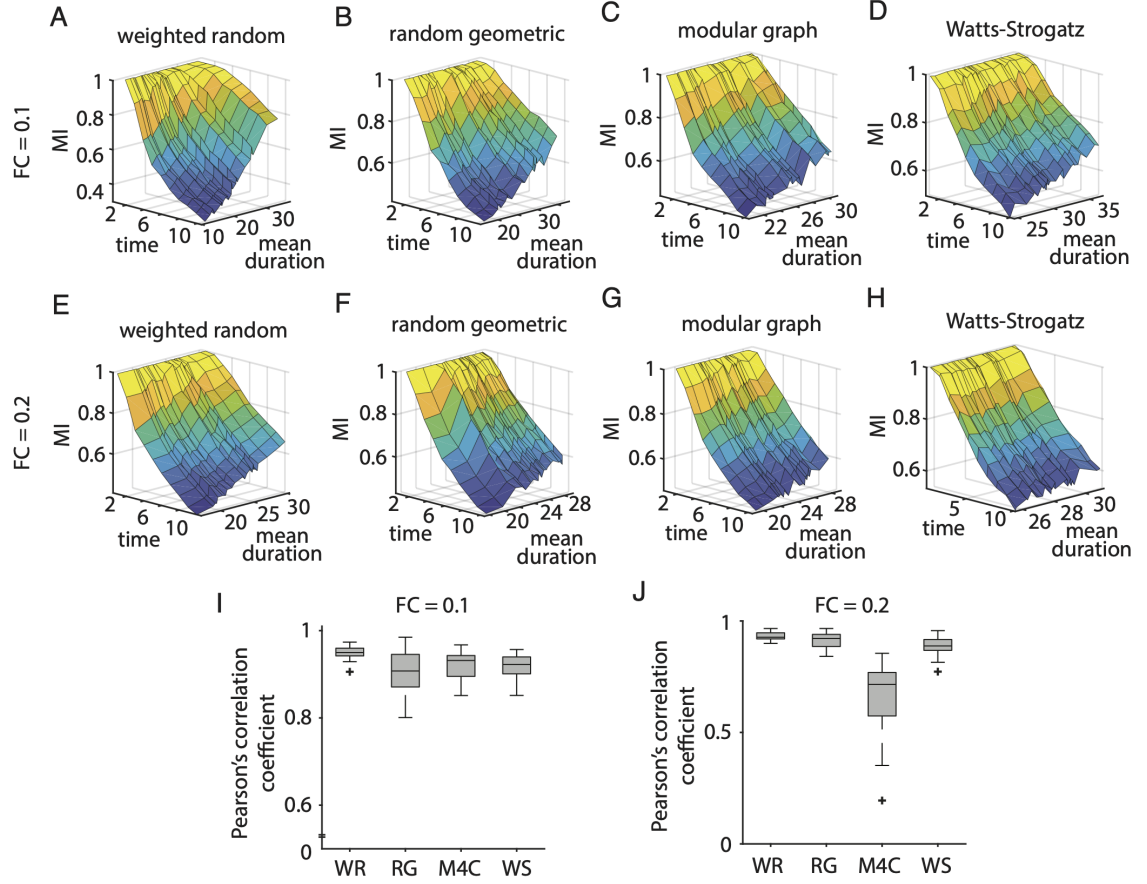


Figure 3.10: **Longer cascade duration allows stimulus recovery.** **A-D** Decay in mutual information (MI) over time. When activity from a stimulus pattern lasts longer, mutual information also persists for longer. Panels show results from four graph types: a weighted random graph, a random geometric graph, a modular graph with 4 communities, and a Watts-Strogatz graph. The networks for these panels have fractional connectivity of 0.1 and a bimodal distribution of weights. **E-H** The same results are presented here as in panels A-D for networks with a fractional connectivity of 0.2. **I** A boxplot of the Pearson correlation coefficients between the linear slope of decay in mutual information over time and the mean cascade duration. The boxplot shows data from 30 instantiations of each graph type, each network containing 100 nodes and characterized by a fractional connectivity of around 0.1. The box-plot elements, center, bottom and top edges, whiskers, “+” symbols, indicate respectively, the median, 25th and 75th percentiles, extremes, and outliers. **J** The same results are presented here as in panel I for networks with fractional connectivity of 0.2.

bin  $t$  and the branching parameter from one time bin  $t$  to the next  $t + 1$ , defined s

$$\sigma_t = \frac{n_{t+1}}{n_t}$$

where  $n_t$  is the number of neurons that are active at  $t$ .

If activity is diluted enough in a network such that collisions can be ignored, then there should be little to no correlation between the number of neurons at  $t$  and the branching parameter  $\sigma_t$  at  $t$ . However, we find a negative correlation between the two variables in the empirical data (Figure 3.11 in this supplementary document). For the 25 recordings, the Spearman's rank correlation coefficient ranges from -0.64 to -0.32 (Figure 3.11B; see example in Figure 4A). When we measure the correlation for cascades of all recordings, the Spearman's rank correlation coefficient is -0.41 with a p-value that cannot be distinguished from 0. Note that the number of neurons active in each bin is much smaller than the system size (min: 98, max: 594, mean: 309, total: 7735). This finding suggests that as more neurons fire, spikes collide and less neurons fire in the next time bin.

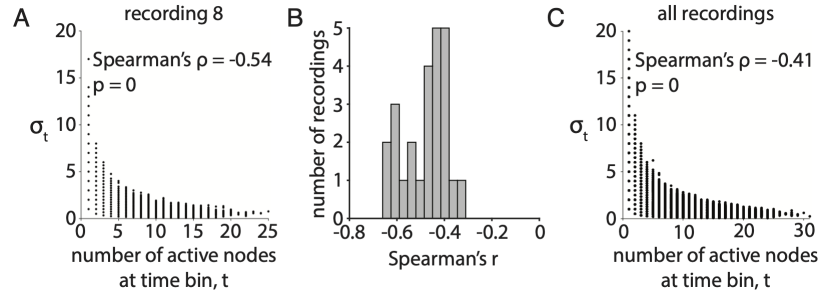


Figure 3.11: **Collisions may occur frequently in living neuronal systems.** **A** The number of active neurons at a time bin  $t$  is negatively correlated with the branching parameter at  $t$  in example recording 8. **B** All 25 recordings display a negative correlation between the number of active neurons and the branching parameter. **C** Cascades for all 25 recordings show a negative correlation between the number of active neurons and the branching parameter.

### 3.8.7. Controllability constrains cascade duration in living neuronal systems

Here, we provide individual plots of finite average controllability and cascade duration for two of the recordings presented in Figure 3.4I. See Figure 3.12. Each point represents a cascade, with the mean finite average controllability of its first  $T$  time bins on the x-axis and its duration on the y-axis.

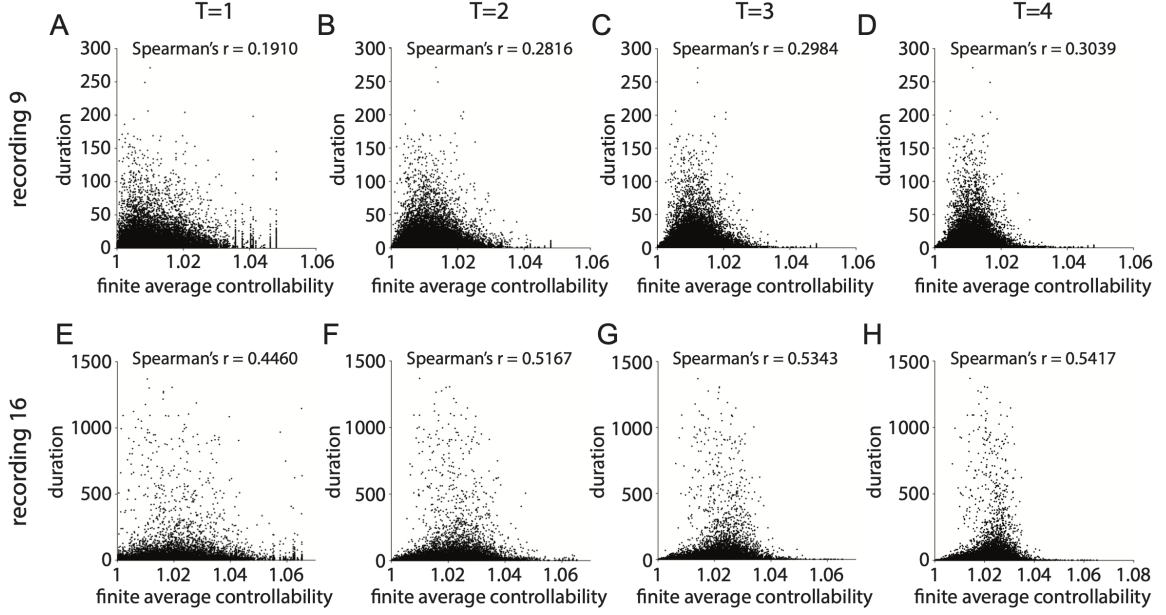


Figure 3.12: **Controllability constrains cascade duration.** **A-D** The finite average controllability of the first  $T$  time bins of a cascade plotted against the duration of the cascade for recording 9. **E-H** The same plots as A-D for recording 16.

### 3.8.8. Multistep regression (MR) estimation of system stability

To understand the constraint of network structure on system stability in addition to the constraint on cascade duration, we estimate the growth parameter of cascading dynamics on different network structures. The growth parameter  $m$  estimates the expected number of spikes at time  $t + 1$  given the number of spikes at time  $t$ . This growth parameter is like the branching parameter of branching processes except for processes with a 1<sup>st</sup> order autoregressive representation (PAR), which approximates our cascading dynamics (see Mathematical Framework). To estimate this parameter even under subsampling, which often leads to a strong overestimation of stability in neural systems, we use multistep regression estimation (Wilting et al., 2018) in simulated cascading dynamics on the synthetic networks used in Figure 3.2.

We find that as the dominant eigenvalue  $\lambda_1$  approach 1 (to a maximum of 0.9973),  $m$  reaches a maximum of 0.9643, and thus, cascading dynamics become *reverberating* and not

*asynchronous-irregular* or *critical* (Figure 3.13) (Wilting et al., 2018). This reverberating behavior was also observed in *in vivo* spike recordings of cat, monkey, and rats (Wilting et al., 2018). However, for networks with a dominant eigenvalue less than approximately 0.7, the MR estimator overestimated the growth parameter of the system. This is likely due to a discrepancy of greater than an order of magnitude between the system size of 256 and a maximum cascade size of less than 10, consequently leading the estimator to assume a smaller system size. In contrast, the dominant eigenvalue still accurately describes the distributions of cascade duration at this range (see Results). Together, these results suggest that (1) cascading systems are *reverberating*, similar to previous findings, and that (2) dynamical systems analysis describes cascading dynamics even in conditions that are not well described by multistep regression.

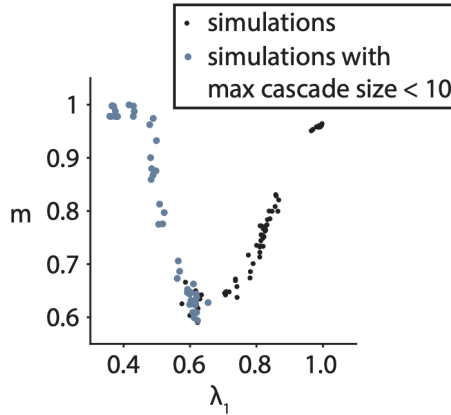


Figure 3.13: **Reverberations in cascading dynamics.** The growth parameter  $m$  decreases initially with the dominant eigenvalue  $\lambda_1$ , but increases with  $\lambda_1$  when the maximum cascade size increases above 10.

### 3.8.9. Comparison of models on empirical distributions of cascade duration

The distributions of cascade duration in empirical MEA recordings and simulations resemble power laws with exponential truncations. Such distributions in granular materials have recently been described by an exponentially truncated power law  $p(x) \sim x^{-\alpha} e^{-x/\tau}$  (Denisov et al., 2016; Murphy et al., 2019), as it can capture distributions that range from exponential to power law. Depending on the value of the parameters (i.e.,  $\alpha = 0$  or  $\tau = \infty$ ), the

truncated power law can become an exponential or a power law, respectively. To determine whether a truncated power law describes the distributions better than either an exponential  $p(x) \sim e^{-x/\tau}$  or a power law  $p(x) \sim x^{-\alpha}$ , we calculated the likelihood ratios between the models (Clauset et al., 2009; Alstott et al., 2014). We found that for all 25 MEA recordings, the truncated power law (TPL) performs better than the exponential (E) or the power law (PL; Figure 3.14).

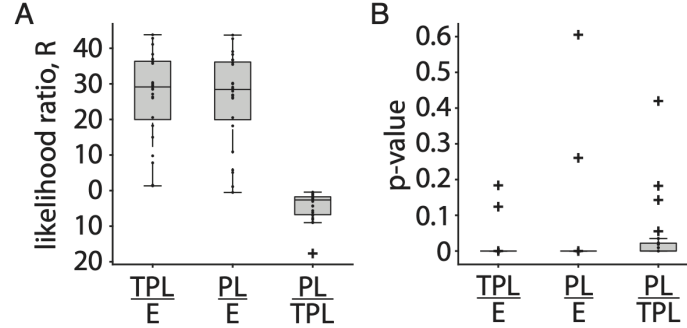


Figure 3.14: **Likelihood ratio test between models on empirical distributions of cascade duration.** **A** Boxplots of the likelihood ratios  $R$  for three models, truncated power law (TPL), power law (PL), and exponential (E), on distributions of cascade duration from 25 MEA recordings. The likelihood ratio is positive if the first distribution (e.g., TPL in TPL/E) is more likely and negative if the second distribution (e.g., E in TPL/E) is more likely. **B** The  $p$ -value of the likelihood ratio test is less than 0.001 for all but two comparisons for TPL/E and for all but two comparisons for PL/E but more variable for comparisons for PL/TPL.

### 3.8.10. Exponent relation test for criticality

The hypothesis of criticality in the brain is still somewhat controversial because a power law of avalanche sizes, which is the most often-used indicator of criticality, can be produced by many mechanisms that are not critical (Beggs and Timme, 2012; Wilting et al., 2019; Touboul et al., 2017). However, one can use the exponent relation (Friedman et al., 2012) as a stricter test for criticality that cannot be produced by models that are not critical, such as molecular chaos models (Touboul et al., 2017). In the exponent relation, the power law exponents,  $\alpha$  and  $\tau$ , of the distributions of cascade duration  $d$  and size  $s$ , respectively, are related to the power law exponent of cascade size given duration  $p(s|d) \sim d^{1/\sigma\nu z}$ , as given

by  $\frac{\alpha-1}{\tau-1} = \frac{1}{\sigma\nu z}$  (Friedman et al., 2012). Recent studies have used this relation to evaluate criticality and identify its presence in their experimental systems (Shew et al., 2015; Ponce-Alvarez et al., 2018; Fontenele et al., 2019). Here, we used the exponent relation to test for criticality in the 25 MEA recordings and found some recordings to be solidly in the critical regime, with a difference between the predicted and fitted exponents  $|\beta_p - \beta_f|$  less than 0.01 (Figure 3.15) (Clauset et al., 2009; Marshall et al., 2016). Other recordings display differences in exponents  $|\beta_p - \beta_f|$  that range from 0.01 to 3.37.

### 3.8.11. Mean absolute errors of VAR models

It is important to evaluate the accuracy of the trained AR model, for example in its ability to predict neural activity of the empirical test data. Here, we report the mean absolute errors (MAE) of the VAR models from the 25 recordings (Figure 3.16). We calculated MAE as

$$\text{MAE} = \frac{1}{N(T-p)} \sum_{i=1}^N \sum_{t=p+1}^T |y_i(t) - \hat{y}_i(t)|,$$

where  $N$  is the number of neurons,  $p$  is the model order,  $T$  is the number of time bins,  $y_i(t)$  is the number of spikes in bin  $t$  for neuron  $i$ , and  $\hat{y}_i(t)$  is the prediction by the VAR model for  $t$  from  $p+1$  to  $T$ . The minimum, mean, and maximum MAE values were 0.0067, 0.0204, and 0.0400, respectively. These findings indicate that the trained AR model provides an accurate prediction of neural activity in the empirical test data.

## 3.9. Supplementary Methods

### 3.9.1. Parameters used in network simulations

Table 3.1 displays parameters of the network simulations reported in the main text.

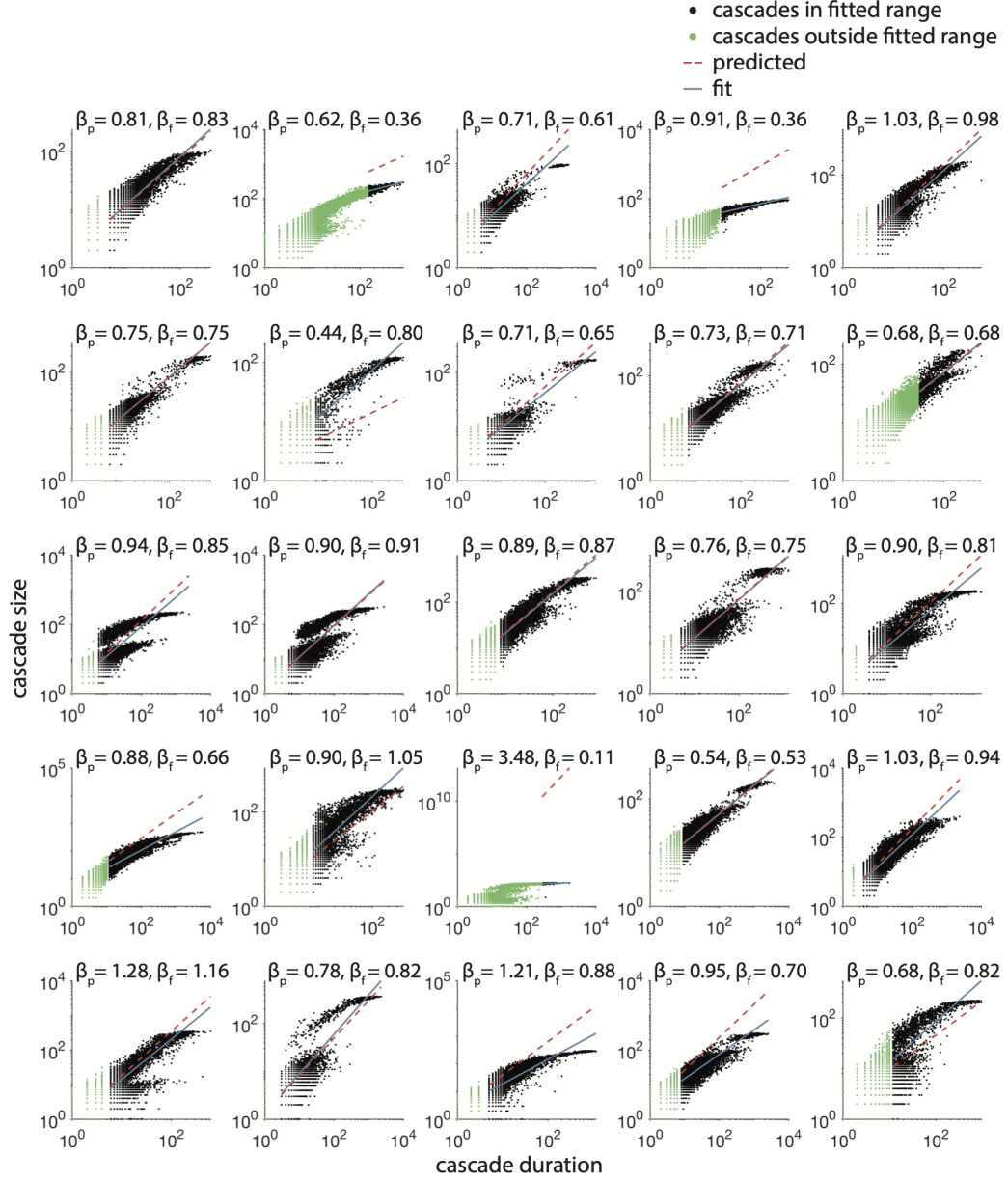


Figure 3.15: **Exponent relations for criticality in MEA recordings.** Cascade size plotted against cascade duration for 25 MEA recordings (recording number from left to right, top to bottom). For some recordings, the predicted exponent  $\beta_p = \frac{\alpha-1}{\tau-1}$  (dashed, red line) matches the fitted exponent  $\beta_f$  (blue line).

### 3.9.2. Predicting cascade dynamics with transition matrix $T$

Given that  $\forall i \in \mathcal{V} : \sum_j a_{ij} \leq 1$  and  $a_{ij} \geq 0$ , We can predict the exact fraction of cascades alive at time  $t$  by computing a state transition matrix from any state  $k$  to any state  $l$ . For

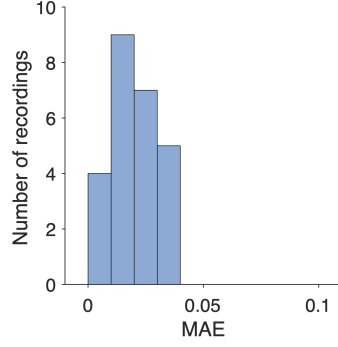


Figure 3.16: Mean absolute errors of VAR models from 25 recordings.

$k, l \in \{1, \dots, n\}$ , the state transition matrix  $T \in \mathbb{R}^{l \times k}$  can be constructed by

$$\begin{aligned}
 P\left(\left[\mathbf{y}(t) = \mathbf{s}^l\right] \mid \left[\mathbf{y}(t-1) = \mathbf{s}^k\right]\right) \\
 &= \prod_{j=1}^n P\left(\left[y_j(t) = s_j^l\right] \mid \left[\mathbf{y}(t-1) = \mathbf{s}^k\right]\right) \\
 &= \prod_{j=1}^n (a_j \mathbf{s}^k \text{ if } s_j^l = 1 \text{ and } 1 - a_j \mathbf{s}^k \text{ if } s_j^l = 0) \\
 &= \prod_j (1 - s_j^l) + (-1)^{s_j^l+1} a_j \mathbf{s}^k.
 \end{aligned}$$

Then, at  $t$  for all  $l$ , the probability of the network being in any state is given by

$$\begin{aligned}
 P(\mathbf{y}(t) = \mathbf{s}^l) &= \sum_{k=1}^n P(\left[\mathbf{y}(t-1) = \mathbf{s}^k\right]) \\
 &\quad P(\left[\mathbf{y}(t) = \mathbf{s}^l\right] \mid \left[\mathbf{y}(t-1) = \mathbf{s}^k\right]).
 \end{aligned}$$



Figures	Number of neurons	Fractional connectivity	Topology	Weighting
Figure 1	10	0.2	WR	UN
Figure 2a-c	10	0.15	WR	UN
Figure 2d-f	256	0.6, 0.1 0.14, 0.18	WR, HM WS, M4C	TG ( $\sigma=0.3$ , 0.4,0.5)
Figure 2 extended	12	0.2	WR, RG, M4C, WS	UN
Figure 3a	3	0.2	acyclic, cyclic	UN
Figure 3b	10	0.33	DAG, WR	UN
Figure 3e-h	2	1.0	cyclic	sweep
Figure 3i-l	4	0.5	cyclic	sweep
Figure 4a-d	100	0.2	WR	UN
Figure 4e-h	100	0.2	WR	BG
Figure 5e-j	100	0.2	WR, RG M4C, WS	BG

Table 3.1: **Network parameters for all simulations.** The graph topologies are weighted random (WR), random geometric (RG), modular with 4 communities (M4C), Watts-Strogatz (WS), and hierarchical modular (HM). The weight distributions are uniform (UN), truncated Gaussian (TG), and bimodal Gaussian (BG).

### 3.10. Supplementary Discussion

#### 3.10.1. Neuronal avalanches versus cascades

Neuronal avalanches are cascades of spontaneous neuronal activity that follow a power law distribution of sizes and durations that is typical of avalanches and other critical systems (Bak et al., 1987; Beggs and Plenz, 2003). Neuronal cascades, however, do not always display critical behavior. While empirical distributions of cascade size seem to follow power laws, empirical distributions of cascade duration, also referred to as life times, display a wide range of power law exponents from -1.0 to -2.6 (Beggs and Plenz, 2003; Petermann et al., 2009; Hahn et al., 2010; Friedman et al., 2012; Poil et al., 2012; Lombardi et al., 2014; Bellay et al., 2015; Shew et al., 2015; Ponce-Alvarez et al., 2018). Moreover, it is clear even without rigorous statistical methods that many empirical distributions of event duration do not follow power laws, but more closely resemble exponential distributions. Given this

mixture of both critical and non-critical propagation, we decided to use the term “neuronal cascades” to refer to both types of activity; in our simulations and analyses we do not assume that all such cascades are critical.

## REFERENCES

- Alstott, Jeff, Ed Bullmore, and Dietmar Plenz (Jan. 2014). “powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions.” In: *PLOS ONE* 9.1, pp. 1–11. DOI: 10.1371/journal.pone.0085777.
- Bak, Per, Chao Tang, and Kurt Wiesenfeld (July 1987). “Self-organized criticality: An explanation of the  $1/f$  noise.” In: *Physical Review Letters* 59.4, pp. 381–384.
- Beggs, John M. and Dietmar Plenz (2003). “Neuronal Avalanches in Neocortical Circuits.” In: *Journal of Neuroscience* 23.35, pp. 11167–11177. DOI: 10.1523/JNEUROSCI.23-35-11167.2003.
- Beggs, John and Nicholas Timme (2012). “Being Critical of Criticality in the Brain.” In: *Frontiers in Physiology* 3, p. 163. DOI: 10.3389/fphys.2012.00163.
- Bellay, Timothy et al. (July 2015). “Irregular spiking of pyramidal neurons organizes as scale-invariant neuronal avalanches in the awake state.” In: *eLife* 4. Ed. by Frances K Skinner, e07224. DOI: 10.7554/eLife.07224.
- Clauset, A., C. Shalizi, and M. Newman (2009). “Power-Law Distributions in Empirical Data.” In: *SIAM Review* 51.4, pp. 661–703. DOI: 10.1137/070710111.
- Denisov, D. V. et al. (Feb. 17, 2016). “Universality of slip avalanches in flowing granular matter.” In: *Nature Communications* 7, 10641 EP.
- Fontenele, Antonio J. et al. (May 2019). “Criticality between Cortical States.” In: *Phys. Rev. Lett.* 122.20, p. 208101. DOI: 10.1103/PhysRevLett.122.208101.
- Friedman, Nir et al. (May 2012). “Universal Critical Dynamics in High Resolution Neuronal Avalanche Data.” In: *Phys. Rev. Lett.* 108.20, p. 208102. DOI: 10.1103/PhysRevLett.108.208102.
- Garlaschelli, Diego (2009). “The weighted random graph model.” In: *New Journal of Physics* 11.7, p. 073005.
- Gu, Shi et al. (2015). “Controllability of structural brain networks.” In: *Nature Communications* 6.1, p. 8414. DOI: 10.1038/ncomms9414.

- Hahn, Gerald et al. (2010). “Neuronal Avalanches in Spontaneous Activity In Vivo.” In: *Journal of Neurophysiology* 104.6, pp. 3312–3322. DOI: 10.1152/jn.00953.2009.
- Lee, D. S. et al. (Mar. 1, 2004). “Branching Process Approach to Avalanche Dynamics on Complex Networks.” In: *Journal of the Korean Physical Society* 44.3, pp. 633–637.
- Lombardi, Fabrizio et al. (2014). “On the temporal organization of neuronal avalanches.” In: *Frontiers in Systems Neuroscience* 8, p. 204. DOI: 10.3389/fnsys.2014.00204.
- Marshall, Najja et al. (2016). “Analysis of Power Laws, Shape Collapses, and Neural Complexity: New Techniques and MATLAB Support via the NCC Toolbox.” In: *Frontiers in Physiology* 7, p. 250. DOI: 10.3389/fphys.2016.00250.
- Murphy, Kieran A., Karin A. Dahmen, and Heinrich M. Jaeger (Jan. 2019). “Transforming Mesoscale Granular Plasticity Through Particle Shape.” In: *Phys. Rev. X* 9.1, p. 011014. DOI: 10.1103/PhysRevX.9.011014.
- Petermann, Thomas et al. (2009). “Spontaneous cortical activity in awake monkeys composed of neuronal avalanches.” In: *Proceedings of the National Academy of Sciences* 106.37, pp. 15921–15926. DOI: 10.1073/pnas.0904089106.
- Poil, Simon-Shlomo et al. (2012). “Critical-State Dynamics of Avalanches and Oscillations Jointly Emerge from Balanced Excitation/Inhibition in Neuronal Networks.” In: *Journal of Neuroscience* 32.29, pp. 9817–9823. DOI: 10.1523/JNEUROSCI.5990-11.2012.
- Ponce-Alvarez, Adrián et al. (Nov. 2018). “Whole-Brain Neuronal Activity Displays Crackling Noise Dynamics.” In: *Neuron*. DOI: 10.1016/j.neuron.2018.10.045.
- Shew, Woodrow L. et al. (2015). “Adaptation to sensory input tunes visual cortex to criticality.” In: *Nature Physics* 11.8, pp. 659–663. DOI: 10.1038/nphys3370.
- Touboul, Jonathan and Alain Destexhe (Jan. 2017). “Power-law statistics and universal scaling in the absence of criticality.” In: *Phys. Rev. E* 95.1, p. 012413. DOI: 10.1103/PhysRevE.95.012413.
- Wilting, J and V Priesemann (2019). “25 years of criticality in neuroscience — established results, open controversies, novel concepts.” In: *Current Opinion in Neurobiology* 58, pp. 105–111. DOI: 10.1016/j.conb.2019.08.002.

Wilting, Jens and Viola Priesemann (2018). “Inferring collective dynamical states from widely unobserved systems.” In: *Nature Communications* 9.1, p. 2325. DOI: 10.1038/s41467-018-04725-4.

## CHAPTER 4

### LOGIC IN CASCADING NEURAL NETWORKS

*This chapter contains work from Ju, H. and Bassett, D.S. (in preparation). "Digital neural logic."*

#### 4.1. Abstract

Neural information processing is critical for cognition. Networks of spiking neurons underlie information processing, and such networks are often modeled as networks of pairwise neural interactions, in which one neuron affects another, independently of or linearly with the influence of other neurons. Recent methods in information theory have begun to quantify the information processed in higher-order neural interactions between more neurons. However, it is still unclear how higher-order neural interactions can be characterized. Here, we formulate triplet neural interactions as probabilistic logic gates to describe neural computations in spontaneous activity in the mouse cortex and the rat hippocampus. Moreover, we find that as neural connections develop *in vitro*, computations become more diverse and that these logic gates follow a characteristic temporal pattern. Taken together, these results demonstrate the utility of studying higher-order interactions to understand neural computation.

#### 4.2. Introduction

Neural computation supports cognition. Recent studies have taken many approaches to quantifying and characterizing neural computation with tools available from a variety of fields, including information theory, dynamical systems, and statistics (Carandini et al., 2012; Lynn and Bassett, 2019; Ju et al., 2020). Information transmission is one of the major areas of interest, by which neurons transmit information to other neurons. Tools from information theory have been critical in quantifying information transmission across neurons (Vicente et al., 2011). Across brain regions, information transmission is studied

as statistical correlations between brain areas (Friston, 1994; Kriegeskorte, 2008). Neural memory is another aspect of neural computation that has been studied extensively. From classical attractor models to many modern methods of neural representation, reservoirs, and high-dimensional coding, studies have described the ways in which neurons can store and transmit information (Rabinovich et al., 2008; Ju et al., 2020; Jaeger, 2001; Stringer et al., 2019).

Neurons, however, cannot simply store and transmit information; they must also provide non-trivial mappings from multiple inputs to outputs. Recent information theoretic measures have further quantified many-to-one mappings of inputs to outputs. Partial information decomposition separates the information from multiple sources that are independent, dependent, and synergistic (Wibral, Priesemann, et al., 2017; Wibral, Finn, et al., 2017). This method is unique in that it measures “synergistic” information, i.e., neural activity that can only be predicted by the combination of activity from other neurons. In modern silicon-based computing, much of the computation is determined by multiple inputs through logic gates and registers (Vahid, 2011).

In this study, we expand upon the quantification of neural computation from measuring an amount of computation to characterizing the dynamics of computation. We characterize synergistic neural computation as a “neural logic gate” and formally define a gate as the probability of a neuron spiking conditional on the spikes of two other neurons at previous time points. We find that neurons do indeed cluster into a characteristic patterns of firing. Moreover, they differ across brain regions and across development *in vitro*. Our findings not only shed light upon the synergistic computations that neurons perform but also open the ways for better characterization of neural computation at the cellular scale, which can later be composed into higher-order computations.

### 4.3. Results

#### 4.3.1. Identifying logic gates

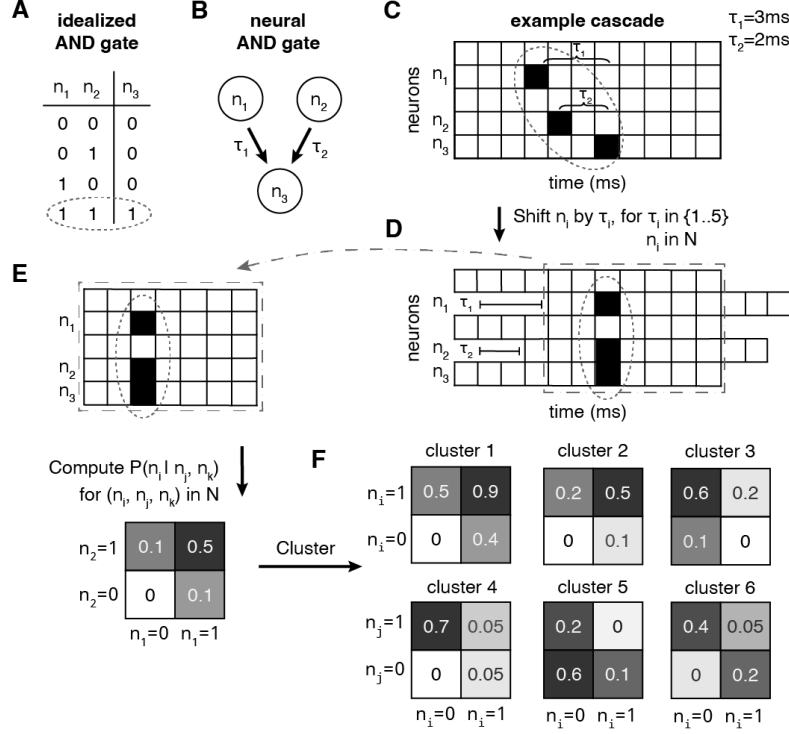


Figure 4.1: **Neural logic gates and how to find them.** **A** An idealized logic gate with two inputs,  $n_1$   $n_2$ , and one output  $n_3$ . **B** A neural logic gate with two inputs,  $n_1$   $n_2$ , each with its own lag  $\tau_1$  and  $\tau_2$ , and one output  $n_3$ . **C** An example of a raster matrix of a cascade. Neurons  $n_1$  and  $n_2$  transmit spikes to neuron  $n_3$  with lags  $\tau_1 = 3\text{ms}$  and  $\tau_2 = 2\text{ms}$ , respectively. **D** To more easily compute the conditional probabilities, shift neurons  $n_1$  and  $n_2$  by lags  $\tau_1$  and  $\tau_2$ , respectively, so that the  $n_1$ ,  $n_2$ , and  $n_3$  are aligned. **E** Compute the conditional probability  $P(n_3 | n_1, n_2)$  for all triplets  $(n_1, n_2, n_3)$  in the set of all neurons  $N$ . To visualize the vectors of conditional probabilities, reshape them as a 2-by-2 matrix, with neurons  $n_1$  and  $n_2$  in each dimension. **F** Cluster the resultant conditional probability vectors with hierarchical agglomerative clustering.

A logic gate is an idealized model of computation that implements a Boolean function. Logic gates are traditionally instantiated on silicon transistors and are deterministic with negligible temporal delays (Figure 4.1A). On neural substrates, neural activity requires axonal and synaptic transmission over millisecond time frames (Figure 4.1B). Thus, we define a two-input “logic gate” as the conditional probability of a target neuron  $t$  firing given two other



source neurons  $s_1$  and  $s_2$  at, respectively, lags of  $\tau_1$  and  $\tau_2$  millisecond time bins before  $t$  firing,  $P(t|s_1, \tau_1, s_2, \tau_2)$ .

To quantify the triplet neural interactions in real neural systems, we clustered triplet conditional probabilities of spontaneous spiking activity. To ensure that we detect neural interdependencies at high temporal resolutions, we first binned the data into 1 millisecond time bins. Then, we partitioned the spike times into “cascades”, or continuously active bins of spikes (Figure 4.1C). By partitioning into cascades, we avoid calculating conditional probabilities between the spontaneous start points and end points of continuous neural activity. Then, we iterate over neurons: one target neuron and two source neurons that precede the target by lags  $\tau$ , ranging from 1 to 5 ms. To align the neurons, we shift the bins for the source neurons  $n_1$  and  $n_2$  by  $\tau_1$  and  $\tau_2$ , respectively (Figure 4.1D). With the shifted cascade matrix, we compute the probability of firing for  $n_3$  conditional on the firings for  $n_1$  and  $n_2$  (Figure 4.1E), from which we form clusters of conditional probabilities (Figure 4.1F).

#### 4.3.2. Model

To test whether our method quantifies logic gates, we first create a model of probabilistic, two-input logic gates (Figure 4.2A). We then create a raster matrix with  $n$  neurons and  $b$  time bins (Figure 4.2B). To model “logic gates” in the raster matrix, we set a high probability ( $P = 0.9$ ) of firing for one neuron based on the states of two other neurons and a predetermined mapping as shown in Figure 4.2A. In this simulation, we map inputs  $n_1$  and  $n_2$  to  $n_3$  as an XOR gate, such that  $n_3$  fires when either of  $n_1$  or  $n_2$  fires, but not both. To simulate the neurons, we spontaneously fire all neurons with a predetermined low firing rate 0.00001, with a slightly higher firing rate 0.001 for the neurons that input to the gates (Figure 4.2B) and with a high firing rate 0.9 for  $n_3$  depending on the states of  $n_1$  and  $n_2$ .

From our simulations, we observed that we can detect arbitrary gates using conditional probabilities. When we clustered the conditional probabilities as explained in the previous section, we obtained an “XOR” cluster that reflects the “XOR” gate in the simulation

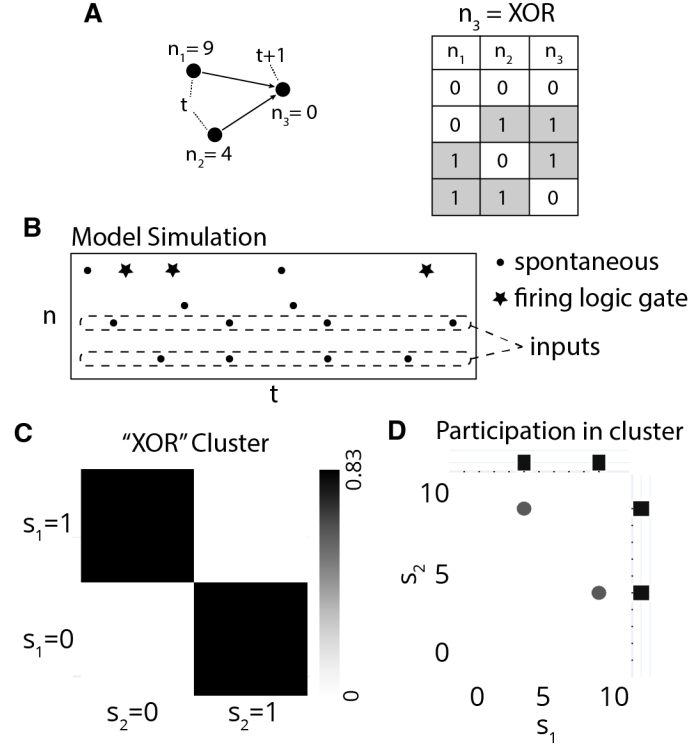


Figure 4.2: **Model of logic gates.** **A** Simple probabilistic model of a two-input “XOR” logic gate. **B** Illustration of model simulation with  $n$  neurons for  $t$  time bins. The simulation spontaneously fires neurons at a low, predetermined firing rate of 0.001 and models the “XOR” gate from panel A. **C** By clustering triplet conditional probabilities, illustrated in Figure 4.1, we detect the “XOR” cluster. The heatmap shows the average probability of a target neuron  $t$  firing, conditional on whether sources  $s_1$  and  $s_2$  have fired. **D** The panel shows which source neurons, out of 10 in the simulation, participated in the cluster in panel C. We can see that only the neurons that we set as the source neurons in panel A participate in this cluster.

(Figure 4.2C). See Figure 4.7 for all clusters. To further verify that the “XOR” cluster was indeed a result of our “XOR” gate, we checked that the source neurons  $s_1$  and  $s_2$  for the “XOR” cluster were 9 and 4, respectively in Figure 4.2D as in Figure 4.2A.

#### 4.3.3. Cortical neurons

To quantify the triplet neural interactions in real neural systems, we clustered conditional probabilities of spontaneous spiking activity of hundreds of neurons in mouse somatosensory cortex slice cultures (Ito et al., 2016). To ensure that we detect neural inter-dependencies

at high temporal resolutions, we first binned the data into 1 millisecond time bins. Then, we partitioned the spike times into “cascades”, or continuously active bins of spikes. By partitioning into cascades, we avoid calculating conditional probabilities between the spontaneous start and ends of continuous neural activity. Then, as explained in previous sections, we clustered the conditional probabilities.

In our analyses, we found that cortical neurons cluster into probabilistic yet distinct logical gates (Figure 4.3A). First, we can qualitatively describe the probabilistic logic gates. In the subsample of 30 neurons in Figure 4.1A, we can see that in cluster 0, the target is more likely to fire when  $s_2 = 1$  but less likely to fire when  $s_1 = 1$  at all, perhaps suggesting that  $s_1$  in this triplet is inhibiting the target while  $s_2$  is exciting the target. We can observe that the neurons that participate in each cluster is non-random (Figure 4.8). In cluster 1, the target is likely to fire when neither of the sources are firing. This is also the largest cluster with the lowest conditional probability, suggesting that this cluster results from spontaneous activity. Clusters 2 and 5 show similar behaviors as in cluster 0. Cluster 3 is interesting because it is an AND gate, i.e., the target fires only when both sources fire, with the highest conditional probability. In cluster 4, the target is more likely to fire when the sources fire and even more when both sources fire, suggesting linear dynamics between the sources and the target.

To test whether the clusters capture statistical dependencies between neurons, we compared clusters detected in the real data to clusters of a null model. We constructed the null model by shuffling the spike times for spikes within each cascade. Thus, we hold constant the firing rates of neurons and cascades statistics (Figure 4.9); the null model only removes the statistical dependence, if there is any, between neurons. We found that that silhouette scores, which measure the quality of clustering, were higher for real data than for null data in 30 subsamples in Dataset 3 (Figure 4.3B). These results suggest that real neurons *in vitro* synergize to produce patterns of higher-order interactions.

Further, we compared logic gates from the real data with different null models to determine

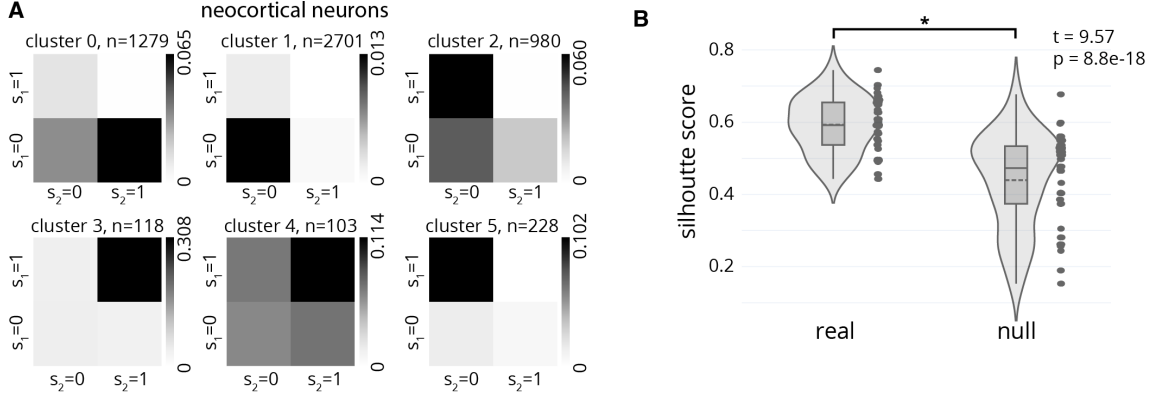


Figure 4.3: **Logical gates in cortical neurons.** **A** Centroids of six clusters of triplet conditional probabilities from dataset 3 from mice somatosensory cortex. The centroids are clustered from a random subsample of 30 neurons from 99 total neurons in the recording; the method of choosing a random subsample was employed for computational tractability. **B** Silhouette scores of 30 random subsamples, each with 30 out of 99 neurons in the recording. Clusters obtained from real data have higher silhouette scores than clusters obtained from time-randomized null models, indicating that real data more clearly form clusters than randomized data.

the dimensions of the data that are relevant for synergistic neural interactions. First, to test whether logic gates require millisecond resolution, we averaged the conditional probabilities across lags from 1 ms to 5 ms between the sources and each target. We found that averaging across 1 to 5 ms lags removed any difference in synergistic interactions between the real and null data (Figure 4.10). Secondly, to test whether cascades are important for computation, we compared real to null data for the entire spike train without first partitioning the spike trains into cascades (Figure 4.11). We found that without first partitioning the spike trains into cascades, the difference in silhouette scores between real and null data disappeared. The results here suggest that neural computation occurs both at the millisecond timescale and is specific to cascades.

#### 4.3.4. Hippocampal neurons

Neurons behave differently based on their underlying network structure, which varies by their location in the brain (Passingham et al., 2002; Bassett et al., 2017; Suárez et al., 2020). Here, we wish to quantify how different brain areas, specifically the neocortex and

the hippocampus, may have similar or varying behaviors in their computation (Timme et al., 2016). To quantify neural computations in the hippocampus, we calculated the centroids of the conditional probabilities from spontaneous spiking activity of neurons in rat hippocampal dissociated cultures. Qualitatively, in Figure 4.4A, neurons are mostly firing when no sources have fired, with 8694 triplets in cluster 3. Moreover, most neurons participate in this cluster (Figure 4.4B). The other clusters are either firing when only one of the sources are firing, as in an “XOR” gate (Figure 4.4A). These logic gates are qualitatively distinct from those found in the neocortex.

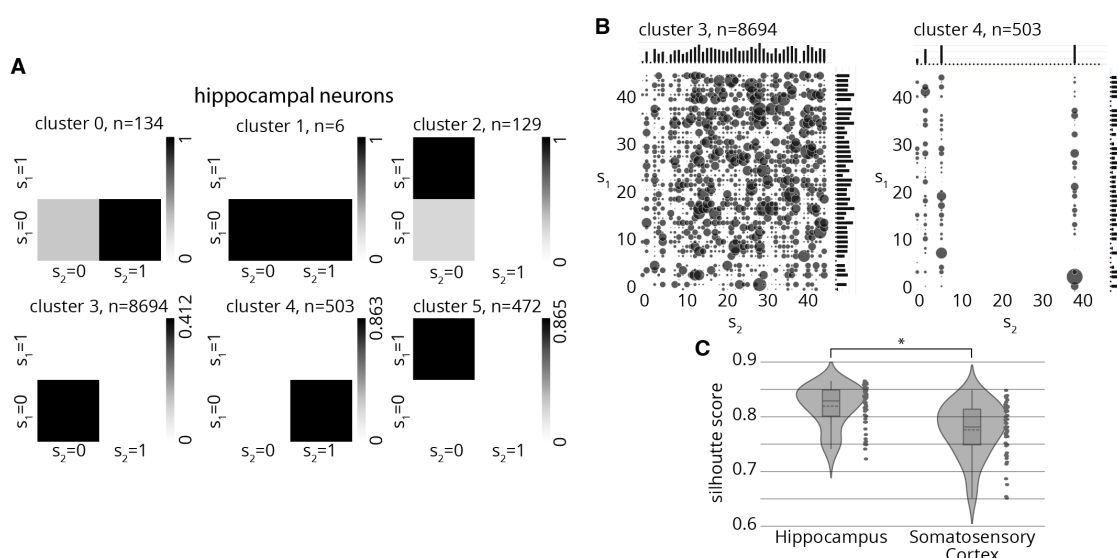


Figure 4.4: **Computations in hippocampal neurons.** **A** Centroids of six clusters of conditional probabilities from slices culture #3 from 30 days *in vitro* from rat hippocampus. **B** Participation of neurons as sources 1 and 2 in cluster 3 and 4, as illustrative examples. Most neurons participate in cluster 3, but not in cluster 4. **C** Conditional probabilities from the hippocampus tend to cluster better than those from the neocortex.

To quantitatively test whether there is a differences between neural computations in the neocortex and the hippocampus, we tested whether triplets from the two areas cluster differentially. First, logic gates in the hippocampus have maximum values that are greater and bimodal in distribution than in the neocortex (Figure 4.4C). Taken together, these results suggest that different brain areas differentially map multiple inputs to outputs, which may underlie more complex neural dynamics.

#### 4.3.5. *In vitro* neural development

While neural computations differ across brain areas, how do they change as neural networks develop? We can quantify changes in neural computation in hippocampal slice cultures across days *in vitro*. While the logic gates look qualitatively similar across days, the distribution of triplets to the centroids appear over time. We found that the gate with the highest conditional probability at  $s_1 = 0$  and  $s_2 = 0$  receives a greater distribution of triplets early on after disassociation (Figure 4.5). However, with more days *in vitro*, more and more triplets are distributed to the other logic gates. These results suggest that as synapses redevelop after disassociation through spontaneous activity, the new connections contribute toward the “XOR-like” gates that we see in the hippocampal slice cultures.

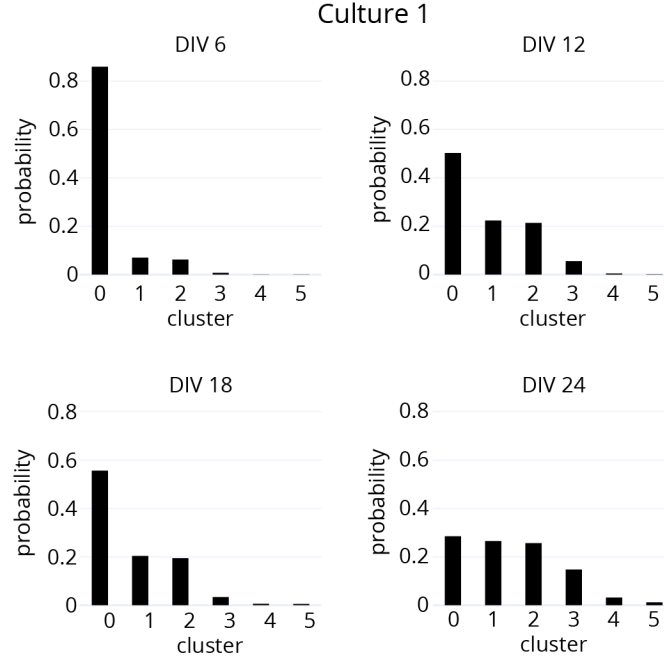


Figure 4.5: ***In vitro* neural development.** The probability distributions to 6 clusters from 6 days *in vitro* (DIV) to 24 DIV. The probability distribution to clusters *evens out* as neurons develop. Clusters have been reordered such that cluster 0 is always the one with the highest probability where  $s_1 = 0$  and  $s_2 = 0$ , as in cluster 3 in Figure 4.4A.

#### 4.3.6. Spatiotemporal patterns

Synergistic neural interactions do not occur in a vacuum but in a sequence of interactions and, in this case, in cascades of spontaneous activity. To quantify the temporal correlations between the occurrences of logic gates, we calculate a probability matrix for the transitions between clusters. As a brief overview, we first identify the most likely gate for each spike by determining the maximum conditional probability that a neuron spiked based on which neurons fired 1 to 5 ms before (Figure 4.6A). Then, we compute the probability of transitions between one cluster to another sequentially across all spikes. We found that in dataset 3 in the cortical slice recordings, certain clusters transition with a high probability to cluster 3 but not clusters 2 and 5 (Figure 4.6B).

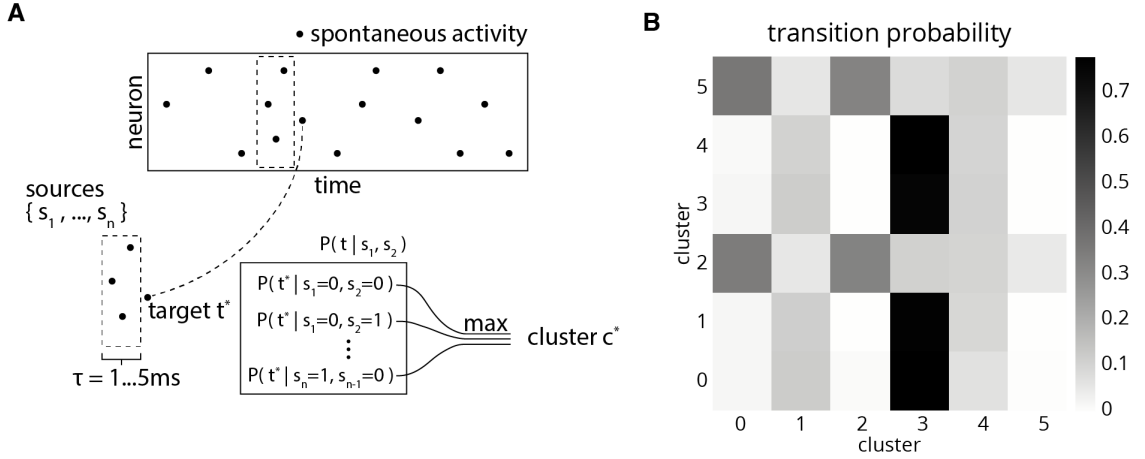


Figure 4.6: **Cluster transition probability.** **A** Illustration of calculating cluster transition probabilities. **B** Transition probability between clusters for dataset 3 in the somatosensory cortex slice recordings.

#### 4.4. Discussion

Here, we expand the scope of study on neural interactions from pairwise statistical dependencies to triplet neural interactions. We characterize triplet interactions by adapting the “logic gate” from digital logic to neural systems as probabilistic, temporal conditional probabilities from two source neurons to one “output” neuron. As we will discuss, this model

is readily expanded to describe  $n$ -to-one neural interactions. We first demonstrate that our formulation forms clusters of triplet neural interactions. Then, we show how such interactions differ across brain regions and across neural development *in vitro*. Finally, we briefly describe temporal patterns in sequences of clusters. Taken together, the results here suggest that quantification of higher-order neural interactions may shed light into neural computations.

#### 4.4.1. triplet interactions

Many modern network models use pairwise interactions (Bassett et al., 2017). Pairwise interactions can explain the behaviors of many complex systems, such as human communication (Lynn, Papadopoulos, et al., 2019). However, higher-order neural interactions are necessary to describe more complex neural dynamics (Wibral, Priesemann, et al., 2017; Wibral, Finn, et al., 2017). With respect to network structure, algebraic topology formulates connections as simplices that encompass higher-order topologies, such as a tetrahedral shape (Ghrist, 2014). With respect to neural dynamics, partial information decomposition is an innovative way to quantify two-input-one-output dependencies and has been expanded to  $n$  inputs (Ince, 2017). As we will discuss in future directions, we hope to expand triplet interactions into higher-order  $n$ -to-one interactions.

#### 4.4.2. Methodological limitations

We use conditional probabilities between two source neurons and one target neuron to describe neural interactions. Conditional probabilities are a simple, non-information theoretic measure of neural interactions. One disadvantage of such a formulation is that conditional probabilities do not discriminate between direct and indirect statistical dependencies between neurons. Information theoretic methods, such as transfer entropy, can subtract redundant information (Vicente et al., 2011). Such an information theoretic approach may also be useful in this context to parse the redundant statistical dependencies across neurons, as does partial information decomposition (Wibral, Priesemann, et al., 2017).



Another methodological limitation comes from the spontaneous and *in vitro* nature of the neural data. Spontaneous activity is limited in its ability to reveal computations that underlie cognition. Animals during behavior receive sensory inputs and exert cognition control that affects neural dynamics (Musall et al., 2019). Though some aspects of neural dynamics adapt to neural inputs (Shew et al., 2015), it remains a question how synergistic neural interactions change as a function of sensory input or behaviors. Thus, it is important to apply the methods demonstrated here to neural activity during animal behaviors to better understand the types of computations that are required for cognition. Alternatively, electrically stimulating neurons via electrodes may allow one to test the conditional probabilities that arise during spontaneous activity.

Another limitation of our data is that the data are recordings of spontaneous activity from slices of mice somatosensory cortex and of rat hippocampus (Timme et al., 2016; Ito et al., 2016). While certain neural dynamics are observed *in vitro*, *in vivo*, and even *ex vivo*, it remains to be seen whether the results demonstrated generalize to *in vivo* measurements (Beggs and Plenz, 2003; Priesemann, Valderrama, et al., 2013; Priesemann, Wibral, et al., 2014; Shew et al., 2015). Moreover, cascades manifest “avalanche”-like behavior whose dynamics are near a critical point of stability (Beggs and Plenz, 2003; Priesemann, Wibral, et al., 2014). Such neural systems are poised near a critical point of stability, in which they have optimal information transmission and memory (Beggs, 2004; Haldeman et al., 2005; Priesemann, Wibral, et al., 2014). Though we demonstrate in this study that avalanche behaviors are important for forming identifiable clusters of high-order neural interactions, it is unknown whether and how avalanche behaviors contribute to or detract from such neural interactions.

#### 4.4.3. Future directions

In this study, we begin to characterize higher-order neural interactions at the scale of triplets: two source neurons and one target neuron. Such a configuration borrows from both digital

logic gates and recent studies in partial information decomposition (Wibral, Priesemann, et al., 2017). However, there may be even higher-order neural interactions. For example,  $n$  neurons may project and converge to a single neuron. In such a case, triplet computations may not capture all such neural interactions. In the future, we hope to apply methods like the information bottleneck to determine the order of neural computation, which will also reduce the computational burden of computing all possible triplets (Tishby et al., 2000).

In addition to higher-order neural interactions, neural computations may be composable into sequential algorithms. In the future, we hope to test whether neural logic gates can chain into longer sequences of computations and to quantify their higher-order characteristics. Modern computer architectures use a sequence of low-level commands on transistors to perform certain computations (Vahid, 2011). We hope to determine whether a parallel exists in neurons on the cellular scale. Interestingly, such composability may accommodate the problem of scale in neural systems: how does one relate neural activity on the cellular level to neural activity in the regional level? By composing descriptions of neural activity from a lower level to a higher level, one can describe coarser levels of neural description with finer resolution data.

## 4.5. Methods

### 4.5.1. Experimental data

For analysis of a real neural system, we use publicly available data derived from slices of spiking neurons in the mouse somatosensory cortex (Ito et al., 2016). The data contain 25 recordings, most of which possess hundreds of neurons (min: 98, max: 594, mean: 309, total: 7735). Each recording is 60 minutes long and was acquired at a sampling rate of 20 kHz. The recordings were acquired from organotypic slice cultures by multielectrode arrays (MEAs), each with 512 electrodes on a 1 mm-by-2 mm area.

For our analysis of neurons in the hippocampus, we use publicly available data derived from

slices of spiking neurons in the rat hippocampus (Timme et al., 2016). Most recordings possess about one hundred neurons (min: 3, max: 142, mean: 91, total: 39,529). Each recording is approximately 60 minutes long and was acquired at a sampling rate of 20 kHz.

#### 4.5.2. Computing logic gates

To compute logic gates, we first define a raster matrix from the spike times which have 0.05 ms temporal resolution. We set the duration of each time bin to 1 ms and bin each spike for each neuron into the matrix. Then, we define a cascade as a set of continuously active times, i.e., at least one neuron is firing in the time bin. By doing the following computations by cascade, we can ignore any effects of spontaneous starts and stops of cascades. Moreover, assuming that spikes have non-zero dependencies among neurons, we remove any cascades that are lower than a predefined number of time bins, 20.

We define logic gates as the conditional probability  $P(t|s_{i,l}, \dots, s_{i,n})$  of whether a “target”  $t$  neuron fires  $t = 1$  conditional on multiple other “source”  $s_{i,l}$  neurons at  $l$  previous time steps for each source. We use lags from 1 ms to 5 ms based on prior work that shows the range of duration for spike transmission (Ito et al., 2016). We also compute logic gates by averaging across to 1 to 5 ms lags to test whether temporal resolutions matter for the millisecond timescale. We test this assumption with a null model. We also test the computations for the entire raster matrix, not by cascades (Figure 4.11).

In this paper, we compute conditional probabilities of triplets in subsets of 30 neurons. Because the complexity of the computations are  $O(n) = n^3$ , any larger subsets were computationally infeasible. Further optimization efforts are required to analyze larger subsets.

After we compute the conditional probabilities  $P(t|s_{1,l_1}, s_{2,l_2})$ , we cluster the probabilities. The conditional probabilities are vectors, with four values for the four combinations of input states  $s_1$  and  $s_2$ , i.e.,  $(s_1, s_2) = (0, 0), (0, 1), (1, 0),$  and  $(1, 1)$ . We cluster these conditional probabilities into a predetermined number of clusters of 6 clusters in the main manuscript.

### 4.5.3. Logic gate model

We simulate probabilistic logic gates to test the computation of logic gates as lagged conditional probabilities. To simulate logic gates, we *a priori* determine probabilities of a target neuron  $t$  conditional on the binary states of source neurons  $s_1$  and  $s_2$ . Then, we simulate neural spikes with a low firing rate of 0.0001 for all neurons and with a higher firing rate of 0.01 for neurons  $s_1$  and  $s_2$ .

### 4.5.4. Null model

To determine what part of the data is important for a particular result, we used a few null models that removes parts of the data. The first null model was a strict null model; it preserved (1) firing rates, (2) spike resolution of 1 ms, and (3) cascade statistics (see SI for cascade statistics). It shuffles the bins for spike times within cascades at 1 ms bins; thus, this model is the strictest null model that tests whether the results actually depend on the statistical dependencies between neurons. The two other null models loosen the preservation of one of three data features. The second null model loosens the spike resolution of 1 ms by averaging the conditional probabilities over 1 to 5 ms. The third null model loosens the cascade statistics by shuffling the spike times for the whole recording, disregarding the cascades.

### 4.5.5. Cluster transition probability

To calculate the transition probability from cluster to cluster, we must first map the spike times to the “logic gate” clusters. To map each spike to “logic gates”, we identify the maximum conditional probability of a target spike,  $P(t|s_1, s_2)$ , with spike times of other neurons 1 to 5 ms prior to the spike time of the target  $t$ . Given the sources and lags that have the maximum conditional probability for a target, we map each spike to a cluster. Then, we calculate the transition probability by calculating a transition probability matrix  $T$  for which the element at row  $i$  and column  $j$  is the probability that a spike for cluster  $i$  is

followed by a spike for cluster  $j$ .

#### **4.6. Code and Data Availability**

The code used to generate this data is fully available at <https://github.com/harangju/neuralogica>. The data used in this study are publicly available at <https://crcns.org/data-sets/hc/hc-8> and <https://crcns.org/data-sets/ssc/ssc-3>.

## REFERENCES

- Bassett, Danielle S and Olaf Sporns (2017). “Network neuroscience.” In: *Nature Neuroscience* 20.3, pp. 353–364. DOI: 10.1038/nm.4502.
- Beggs, J. M. (2004). “Neuronal Avalanches Are Diverse and Precise Activity Patterns That Are Stable for Many Hours in Cortical Slice Cultures.” In: *Journal of Neuroscience* 24.22, pp. 5216–5229. DOI: 10.1523/JNEUROSCI.0540-04.2004.
- Beggs, John M. and Dietmar Plenz (2003). “Neuronal Avalanches in Neocortical Circuits.” In: *Journal of Neuroscience* 23.35, pp. 11167–11177. DOI: 10.1523/JNEUROSCI.23-35-11167.2003.
- Carandini, Matteo and David J. Heeger (Jan. 2012). “Normalization as a canonical neural computation.” In: *Nature Reviews Neuroscience* 13.1, pp. 51–62. DOI: 10.1038/nrn3136.
- Friston, Karl J. (1994). “Functional and effective connectivity in neuroimaging: A synthesis.” In: *Human Brain Mapping* 2.1, pp. 56–78. DOI: 10.1002/hbm.460020107.
- Ghrist, Robert (2014). *Elementary applied topology: edition 1.0*. s. l.: Createspace. ISBN: 978-1-5028-8085-7.
- Haldeman, Clayton and John M. Beggs (Feb. 2005). “Critical Branching Captures Activity in Living Neural Networks and Maximizes the Number of Metastable States.” In: *Phys. Rev. Lett.* 94.5, p. 058101. DOI: 10.1103/PhysRevLett.94.058101.
- Ince, Robin A. A. (2017). “The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal.” In: DOI: 10.48550/ARXIV.1702.01591.
- Ito, Shinya et al. (2016). *Spontaneous spiking activity of hundreds of neurons in mouse somatosensory cortex slice cultures recorded using a dense 512 electrode array*. DOI: 10.6080/K07D2S2F.
- Jaeger, Herbert (2001). “The "echo state" approach to analysing and training recurrent neural networks.” In:

- Ju, Harang and Danielle S. Bassett (June 2020). “Dynamic representations in networked neural systems.” In: *Nature Neuroscience*. DOI: 10.1038/s41593-020-0653-3.
- Kriegeskorte, Nikolaus (2008). “Representational similarity analysis – connecting the branches of systems neuroscience.” In: *Frontiers in Systems Neuroscience*. DOI: 10.3389/neuro.06.004.2008.
- Lynn, Christopher W. and Danielle S. Bassett (May 2019). “The physics of brain network structure, function and control.” In: *Nature Reviews Physics* 1.5, pp. 318–332. DOI: 10.1038/s42254-019-0040-8.
- Lynn, Christopher W., Lia Papadopoulos, et al. (Feb. 2019). “Surges of Collective Human Activity Emerge from Simple Pairwise Correlations.” In: *Physical Review X* 9.1, p. 011022. DOI: 10.1103/PhysRevX.9.011022.
- Musall, Simon et al. (2019). “Single-trial neural dynamics are dominated by richly varied movements.” In: *Nature Neuroscience* 22.10, pp. 1677–1686. DOI: 10.1038/s41593-019-0502-4.
- Passingham, Richard E., Klaas E. Stephan, and Rolf Kötter (Aug. 2002). “The anatomical basis of functional localization in the cortex.” In: *Nature Reviews Neuroscience* 3.8, pp. 606–616. DOI: 10.1038/nrn893.
- Priesemann, Viola, Mario Valderrama, et al. (Mar. 2013). “Neuronal Avalanches Differ from Wakefulness to Deep Sleep – Evidence from Intracranial Depth Recordings in Humans.” In: *PLoS Computational Biology* 9.3. Ed. by Olaf Sporns, e1002985. DOI: 10.1371/journal.pcbi.1002985.
- Priesemann, Viola, Michael Wibral, et al. (2014). “Spike avalanches in vivo suggest a driven, slightly subcritical brain state.” In: *Frontiers in Systems Neuroscience* 8, p. 108. DOI: 10.3389/fnsys.2014.00108.
- Rabinovich, Misha, Ramon Huerta, and Gilles Laurent (July 2008). “Transient Dynamics for Neural Processing.” In: *Science* 321.5885, pp. 48–50. DOI: 10.1126/science.1155564.
- Shew, Woodrow L. et al. (2015). “Adaptation to sensory input tunes visual cortex to criticality.” In: *Nature Physics* 11.8, pp. 659–663. DOI: 10.1038/nphys3370.

- Stringer, Carsen et al. (2019). “High-dimensional geometry of population responses in visual cortex.” In: *Nature* 571.7765, pp. 361–365. DOI: 10.1038/s41586-019-1346-5.
- Suárez, Laura E. et al. (Apr. 2020). “Linking Structure and Function in Macroscale Brain Networks.” In: *Trends in Cognitive Sciences* 24.4, pp. 302–315. DOI: 10.1016/j.tics.2020.01.008.
- Timme, Nicholas M. et al. (2016). *Spontaneous spiking activity of thousands of neurons in rat hippocampal dissociated cultures*. DOI: 10.6080/K0PC308P.
- Tishby, Naftali, Fernando C. Pereira, and William Bialek (2000). “The information bottleneck method.” In: DOI: 10.48550/ARXIV.PHYSICS/0004057.
- Vahid, Frank (2011). *Digital design, with RTL design, VHDL, and Verilog*. 2nd ed. Hoboken, NJ: Wiley. ISBN: 978-0-470-53108-2.
- Vicente, Raul et al. (Feb. 2011). “Transfer entropy—a model-free measure of effective connectivity for the neurosciences.” In: *Journal of Computational Neuroscience* 30.1, pp. 45–67. DOI: 10.1007/s10827-010-0262-3.
- Wibral, Michael, Conor Finn, et al. (2017). “Quantifying Information Modification in Developing Neural Networks via Partial Information Decomposition.” In: *Entropy* 19.9, p. 494. DOI: 10.3390/e19090494.
- Wibral, Michael, Viola Priesemann, et al. (Mar. 2017). “Partial information decomposition as a unified approach to the specification of neural goal functions.” In: *Brain and Cognition* 112, pp. 25–38. DOI: 10.1016/j.bandc.2015.09.004.



# Supplementary Information

## 4.7. Supplementary Results

The following are supplementary figures to Chapter 4.

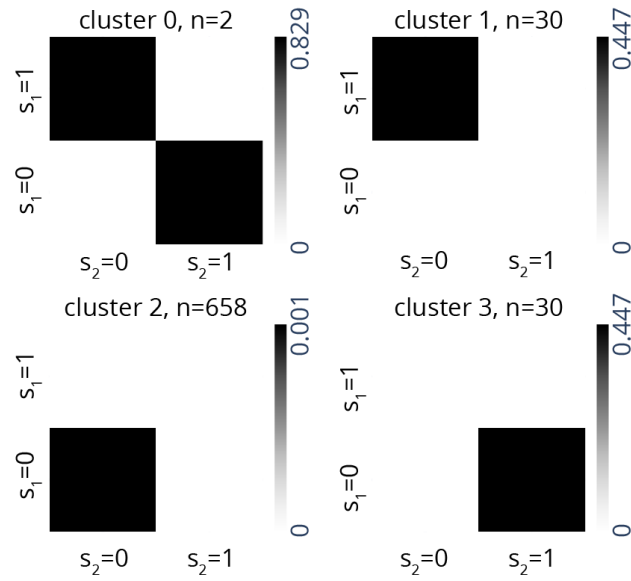


Figure 4.7: **All cluster of the “XOR” model** in the simulation shown in Figure 4.2.

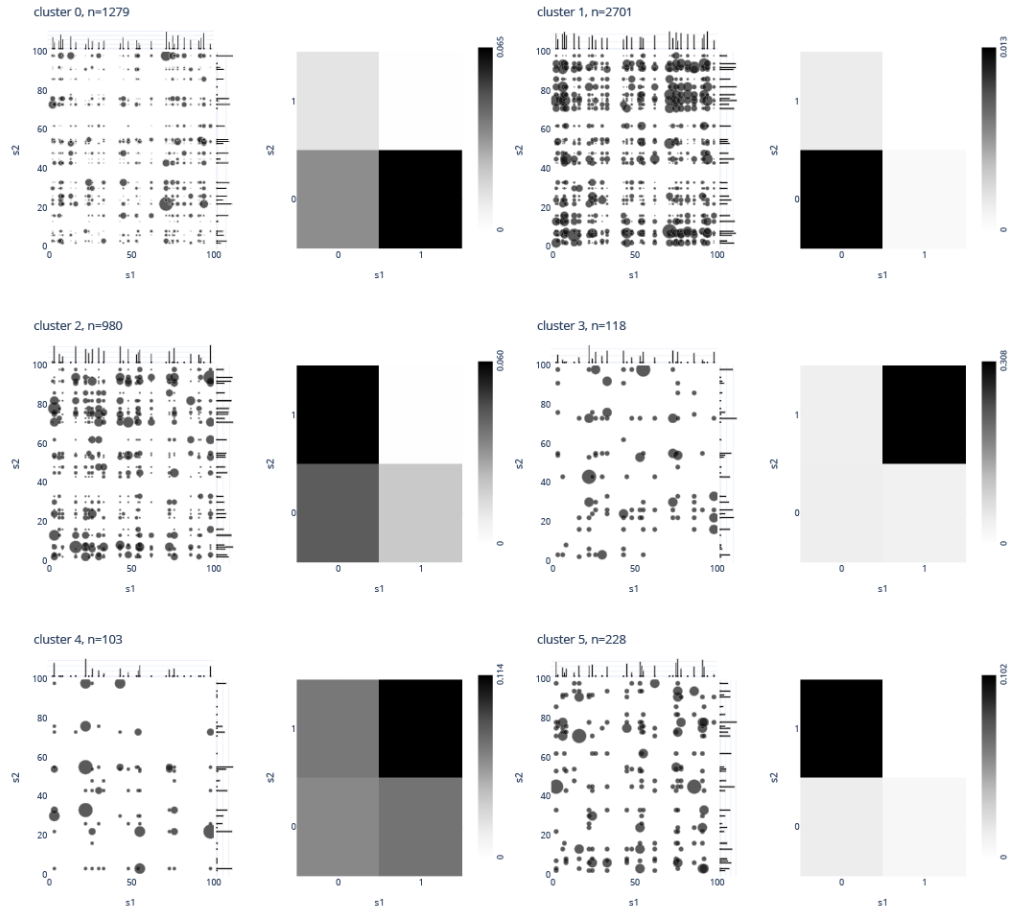


Figure 4.8: **Participation of neurons in clusters.** Each panel shows, on the left, the relative number of times that a neuron participates in the cluster and, on the right, the centroid for the logic gate for the cluster.

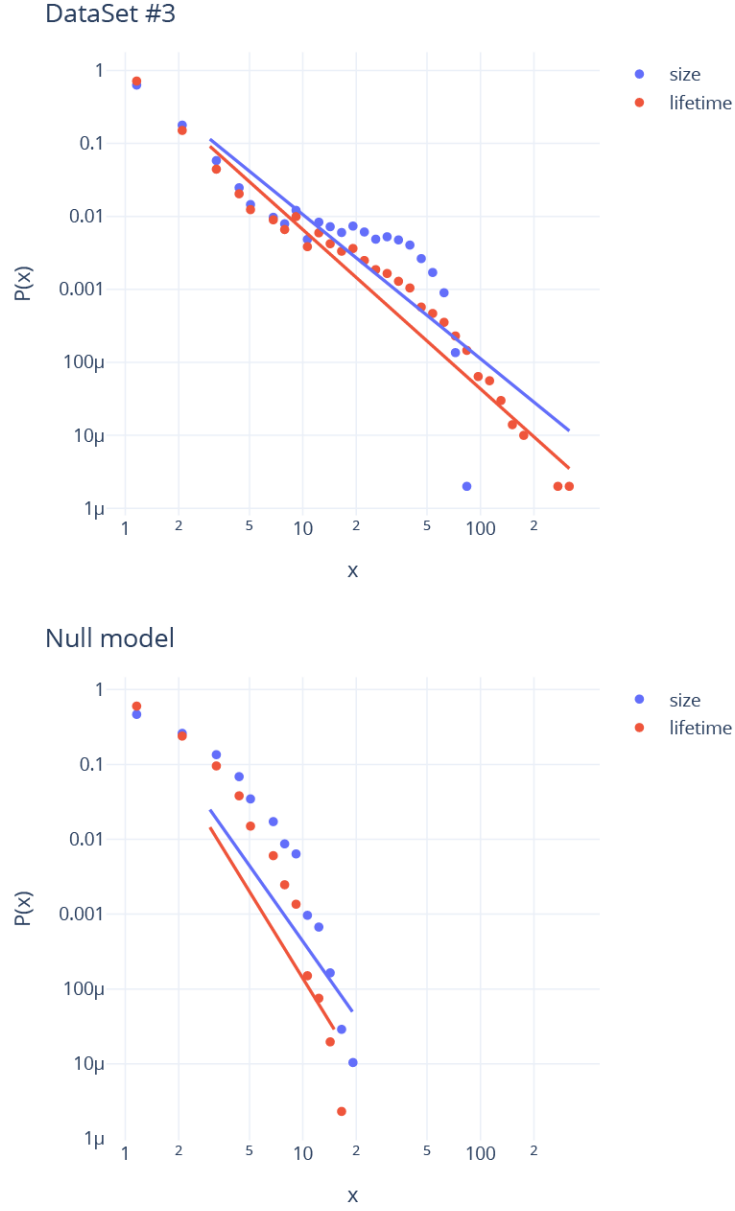


Figure 4.9: **Cascade statistics.** The top panel shows the probability distributions of the size (blue) and lifetime (red) of cascades with the constants 1.97 and 2.18, respectively, for the power law  $p(x) \sim x^{-\alpha}$ . The bottom panel shows the probability distributions of the size (blue) and lifetime (red) of cascades with the power law constants 3.36 and 3.85, respectively. Power law constants greater than 2 have an expected value and thus are not scale-free.

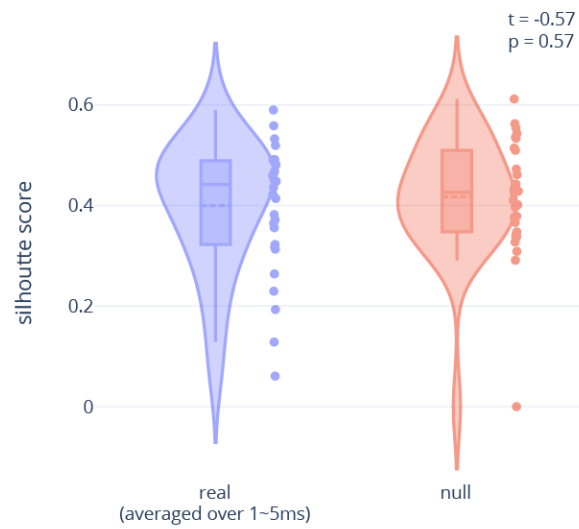


Figure 4.10: **Neural logic gates require spiketimes with millisecond resolution.** The silhouette scores for real data with lags between sources and targets averaged over 1 to 5 ms are not significantly different from those for null data whose spiketimes have been shuffled.

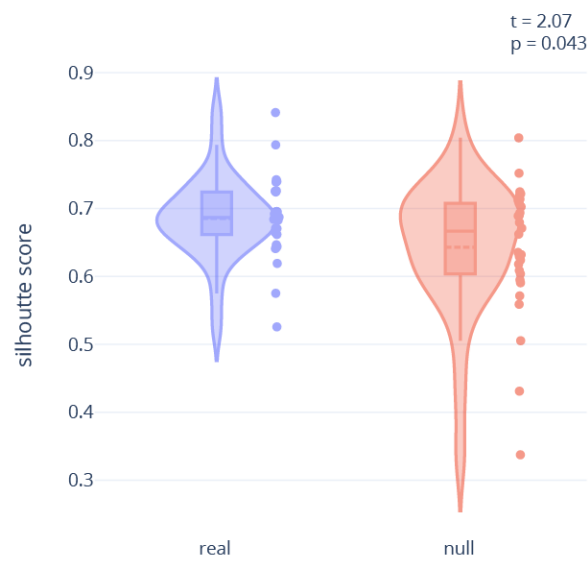


Figure 4.11: **Neural logic gates require cascades.** The silhouette scores for real data whose spiketimes are not grouped by cascades are not significantly different from those for null data whose spiketimes have been shuffled.

## CHAPTER 5

### GROWTH IN CONCEPT NETWORKS

*This chapter contains work from Ju, H., Zhou, D., Blevins, A.S., Lydon-Staley, D.M., Kaplan, J., Tuma, J.R., and Bassett, D.S. (2020). “The network structure of scientific revolutions.” arXiv:2010.08381*

#### 5.1. Abstract

Philosophers of science have long questioned how collective scientific knowledge grows. Although disparate answers have been posited, empirical validation has been challenging due to limitations in collecting and systematizing large historical records. Here, we introduce new methods to analyze scientific knowledge formulated as a growing network of articles on Wikipedia and their hyperlinks. We demonstrate that in Wikipedia, concept networks in subdisciplines of science do not grow by expanding from their central core to reach an ancillary periphery. Instead, science concept networks in Wikipedia grow by creating and filling knowledge gaps. Notably, the process of gap formation and closure may be valued by the scientific community, as evidenced by the fact that it produces discoveries that are more frequently awarded Nobel prizes than other processes. To determine whether and how the gap process is interrupted by paradigm shifts, we operationalize a paradigm as a particular subdivision of scientific concepts into network modules. Hence, paradigm shifts are reconfigurations of those modules. The approach allows us to identify a temporal signature in structural stability across scientific subjects in Wikipedia. In a network formulation of scientific discovery, our findings suggest that data-driven conditions underlying scientific breakthroughs depend as much on exploring uncharted gaps as on exploiting existing disciplines and support policies that encourage new interdisciplinary research.

## 5.2. Significance Statement

Philosophers of science question the way science works and why. For millennia, they have posited mechanisms and offered explanations. Several recent and particularly compelling theories have been difficult to validate, in large part due to challenges in collecting and systemizing large historical records. Fortunately, Wikipedia—the largest online encyclopedia—contains millions of articles that not only form hyperlinked networks of concepts but also include a history of when a concept was discovered. We use this resource to formulate the process of science in terms of knowledge growth, or—more precisely—the growth of concept networks. Across scientific subjects, we demonstrate that networks in Wikipedia, as an important and interesting case, grow both by expanding frontiers and by filling knowledge gaps, striking a balance between bubbling out and bubbling in.

## 5.3. Introduction

In the philosophy of science, thinkers from disparate eras have attempted to reason about the various processes underlying scientific progress. Some of the most compelling theories have come from the last 50 years. For example, in 1959 Karl Popper described the development of scientific ideas as a sequence in which previous theories are falsified (Popper, 2008). In 1962, Kuhn suggested instead that progress was best described as periods of normal science, in which researchers “solved puzzles” within a paradigm, separated from one another by paradigm shifts that overturn the existing paradigm (Kuhn and Hacking, 2012). In 1970, Lakatos balanced the two theories by suggesting that science progresses according to a research programme in which knowledge expands from a common core set of theoretical commitments and practices (Lakatos, 1978). In 1975, Feyerabend differed from the thinkers who had come before by discounting any single mechanism for scientific progress (Feyerabend, 2010). Even more recently, the field of science of science has begun to use a more quantitative approach to probe the cultural, societal, institutional, and personal conditions that support (or do not support) scientific discovery, dissemination, and impact

(Astegiano et al., 2019; Clauset, Larremore, et al., 2017; Zeng et al., 2017; Helmer et al., 2017; Fortunato et al., 2018; Nagaraj et al., 2020; Robinson-Garcia et al., 2020; Wang et al., 2021).

Despite the variety of theories regarding scientific progress, such as cultural accounts of scientific practice (Chemla et al., 2017), many suggest a dependence of new discoveries on the existing body of knowledge. Newton writes in 1675, “If I have seen further, it is by standing on the shoulders of giants” (Newton, 1675). Indeed, discoveries, including calculus, are often multiples, discovered independently and contemporaneously by several scholars, sometimes quite geographically separated (Merton, 1974). These observations prompt the question of how an existing body of knowledge influences the discovery of new scientific knowledge.

Here, we expand upon the concept of a body of knowledge, not as amorphous, but as comprised of distinct relationships between concepts. We formalize this structured body of knowledge as a concept network whose nodes represent concepts and whose edges represent inter-concept relations (Siew et al., 2019). Concept networks have proven to be powerful tools for probing questions about topological structure and the exploration of knowledge (Christianson et al., 2020; Lydon-Staley et al., 2021). To begin to study the process of science, we build growing concept networks from Wikipedia, a free online encyclopedia (Figure 5.1A; Section 5.7). Each Wikipedia article explains a concept and contains hyperlinks to other Wikipedia articles, thereby intuitively forming a directed network of concepts. To formally represent this network, we treat each article as a node. Further, we treat each hyperlink from an article’s lead section (i.e., the introduction) as a directed edge from the hyperlinked article to the hyperlinking article. We weight the directed edges according to the similarity between the two nodes or concepts. More specifically, we calculate the cosine similarity between two articles’ vector representations, which are derived from a term frequency-inverse document frequency (*tf-idf*) encoding. Focusing on the article’s history and lead sections, we parse the earliest year associated with the concept’s discovery and



[illegible]

144

In this study, we first develop methods to operationalize the theories of the philosophers Kuhn, Feyerabend, and Lakatos as falsifiable hypotheses using network theory. We then test each hypothesis—from which we infer the theories that are most supported by the data—in Wikipedia as an important and interesting case. To operationalize and test these hypotheses, we use cutting-edge methods in network science, algebraic topology, and control theory. Here, we find that science concept networks grow by creating and filling knowledge gaps, a process which is valued by the scientific community. Moreover, we operationalize a paradigm as a particular subdivision of scientific concepts into network modules and find that paradigm shifts vary in their magnitude in a similar temporal signature across scientific subjects. Our findings not only shed light upon the large-scale, empirical evidence of classical philosophies of science but also demonstrate operationalizations of qualitative social phenomena that may be readily applied in other contexts.

#### 5.4. Results

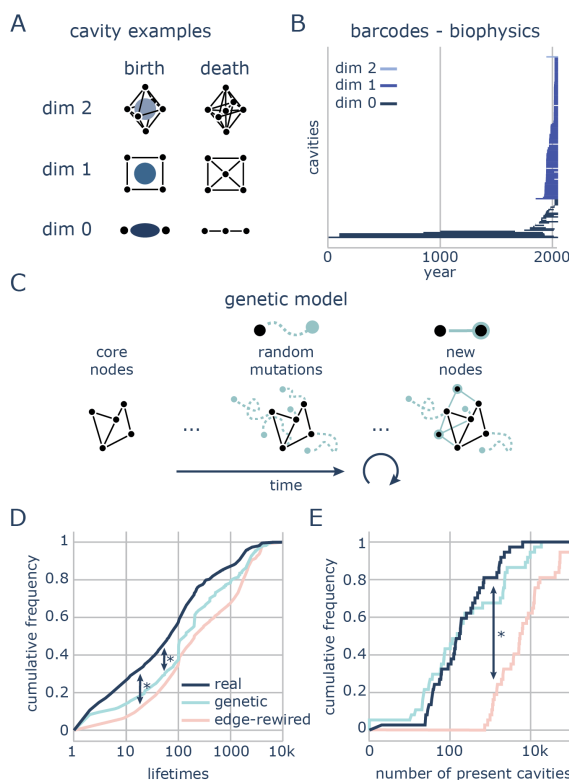
Given the notions of characteristic patterns in discovery by Kuhn and Lakatos, or the lack thereof by Feyerabend, we first quantified the structure of concept networks using computational tools and concepts from network science. As an initial test of Kuhn’s hypothesis of normal science, we measured whether concepts form clusters, in which scientists “solve puzzles” within an existing paradigm. At the node level, we operationalized this idea using the clustering coefficient, which quantifies the local density of connections (Fagiolo, 2007); at the network level, we operationalized this idea using modularity, which quantifies the degree to which groups of nodes are densely connected within distinct modules (Clauset, Newman, et al., 2004). Additionally, we used a core-periphery measure (Borgatti et al., 2000) to assess whether networks form a “common core” as suggested by Lakatos’s research programme, where core nodes are densely connected to each other, and where peripheral nodes are loosely connected to the core. We compared these measures in real networks to those in null networks that destroy existing topology through random rewiring. Randomized null networks may be one way to, thereby operationalize Feyerabend’s hypothesis

of a lack of characteristic pattern to discovery. We found that real networks have greater clustering coefficients, modularity, and coreness than those in the edge-rewired networks (Figure 5.1B). These findings underscore the existence of nontrivial topological structure in concept networks, whereby concepts can either be core to the field or peripheral to the field, and are clustered within modules (akin to subfields).

How might the structure of concept networks arise through the course of history? Might core concepts emerge first, followed by peripheral concepts, in a cumulative growth pattern in which we build deeply upon foundational early discoveries in the field? Or might the “foundational” concepts change as the field grows and changes, such that new discoveries can sometimes replace earlier discoveries as more central to the field? To address this question, we compared the birth year of core nodes to the birth year of neighboring peripheral nodes. We observed that there is no clear lead-lag relationship between a core node and its neighboring peripheral nodes (Figure 5.1C). That is, core nodes do not necessarily precede neighboring peripheral nodes in their discovery, suggesting that the discoveries viewed as “central” to a field can change throughout the field’s lifetime. Notably, there is similarly no clear lead-lag relationship between core and periphery nodes within specific modules (Figure 5.5). Interestingly, a few core nodes are born consistently earlier than most peripheral nodes. These nodes, such as the node “Hydrology” in the subject “Earth Science”, may serve as concepts that are central to the subject as a whole and comprise a “hard inner” part of the core (Figure 5.6). Taken together, these results point to both an outward expansion and an inward exploration of concepts, in which the core of a research programme is often updated by new discoveries that are influenced by discoveries that occur in the periphery.

How then does a body of knowledge grow? Thus far, we understand that real concept networks are highly clustered and display processes of both inward and outward growth. Might concept networks fill gaps in knowledge (Fontana et al., 2020) in a manner that is conceptually akin to Kuhn’s puzzle-solving normal science? To test this possibility, we formalize knowledge gaps in the language of algebraic topology (Hatcher, 2002) and assess

the relevance of gaps to discovery. Specifically and in this mathematical parlance, a gap corresponds to a topological cavity in any dimension  $n$  (Figure 5.2A-B; Section 5.7) (Sizemore et al., 2018). To detect gaps within the growing concept network, we use persistent homology, which chronicles the birth, evolution, and collapse of topological cavities across a growth process (Zomorodian et al., 2005). In general, a cavity in a Wikipedia network is a set of articles (i) that form a single connected component, but (ii) no article in the set connects all articles of the cavity.



**Figure 5.2: Real concept networks maintain shorter and fewer gaps.** **A** Examples of cavities in 0, 1, and 2 dimensions. Cavities are born when a new, added node creates a topological gap; cavities die when the gap is closed by a new node and its edges tessellate the gap. **B** Barcode for the biophysics network. The left and right points of each bar are the birth and death times of a persistent cavity. **C** Illustration of our genetic model for concept network growth and evolution. **D** Real networks have knowledge gaps for shorter duration than either randomly rewired ( $KS = 0.24$ ,  $p = 1.4 \times 10^{-188}$ ) or genetic null model ( $KS = 0.20$ ,  $p = 1.3 \times 10^{-15}$ ) networks. Starred arrows indicate significantly different distributions. **E** Real networks have fewer knowledge gaps that are currently present (i.e., that have yet to die) than random networks ( $KS = 0.81$ ,  $p = 2.1 \times 10^{-12}$ ).

After computing the persistent homology of real networks (Figure 5.2B), we next wished to determine whether the observed gap formation and closure was similar to (or different from) the same processes expected in network null models. We consider two network null models. The first is a randomly rewired null model, in which the edges are relocated uniformly at random. The second is a genetic null model that explicitly examines the process of scientific discovery. The genetic null model simulates the process by which scientists learn about existing concepts and then slowly mutates those concepts to create new ones. Importantly—and unlike true science—the genetic null model contains no preference for how new concepts are mutated, for example as could be operationalized by an objective or fitness function (Figure 5.2C). The model is initialized as a subset of a real network, containing “core nodes” that were born before a predetermined year (BC 500 in our simulations). For each node, we iteratively mutate a *tf-idf* vector representation of the node’s Wikipedia article. We then create a new node from the mutated vector and connect it to similar nodes in the network (Section 5.7; Figure 5.7). This process continues until the number of nodes in the simulated network is equal to that of the true network.

At this point, the topology of the real, rewired, and genetic null model networks can be compared. We find that for persistent cavities that have already died by  $t_{max}$ , those in real networks have significantly shorter lifetimes than those in either randomly rewired networks (Kolmogorov-Smirnov statistic,  $KS = 0.24$ ,  $p = 1.4 \times 10^{-188}$ ) or genetic null model networks ( $KS = 0.20$ ,  $p = 1.3 \times 10^{-15}$ ) (Figure 5.2D). Further, real networks have significantly fewer persistent cavities that are still present at  $t_{max}$  than randomly rewired (but not genetic) networks ( $KS = 0.81$ ,  $p = 2.1 \times 10^{-12}$ ; Figure 5.2E). Collectively, these results support our hypothesis that real concept networks fill gaps more quickly than null networks and leave fewer gaps alive at the present day, consistent with the predictions of Kuhn’s puzzle-solving normal science.

Whereas the first part of Kuhn’s theory regards puzzle-solving, the second part regards paradigm shifts that radically change the way scientists view concepts. In our network

formulation, we operationalized paradigms as temporally varying modular structure, where an “incommensurable” paradigm shift would drastically change the membership of nodes to modules. Accordingly, we built a multilayer network with each layer containing the concept network at each year (Bianconi, 2018) and used multilayer community detection to identify modules at each year and understand how those modules change in constitution over years (Mucha et al., 2010). To summarize these data, we followed prior work to calculate the number of times that a node changes its membership to a module for each year (Figure 5.3A) (Killick, Fearnhead, et al., 2012). Then, we performed change point detection on the number of changes observed to identify epochs of stability in module membership (Figure 5.3B; Section 5.7) (Killick and Eckley, 2014). We found that after an initial, short epoch of little change, the network enters an epoch of moderate and lasting change (Figure 5.3C). Then, the network enters a short epoch of much greater change, after which the network stabilizes in the last epoch. This signature is not observed in randomly rewired networks (Figure 5.8). The data demonstrate that shifts in the structure of concepts occur not as abrupt Kuhnian reconfigurations but as gradual Lakatosian modifications (Daston, 2016). Moreover, the findings demonstrate that subjects (or fields) display a shared signature of structural stability and instability across time.

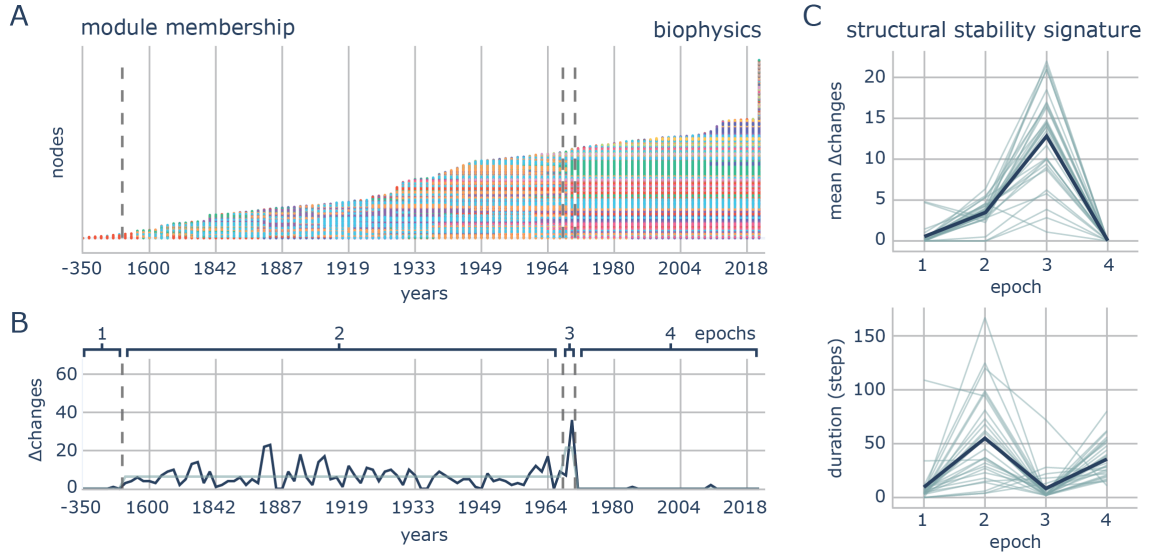


Figure 5.3: **Concept networks undergo a signature pattern in structural stability.** **A** Module membership (colored) of nodes across years for the *biophysics* network. **B** The number of changes in module membership for the *biophysics* network. Dashed lines are changepoints in epochs, and the teal line is the mean for each epoch. **C** The mean number of changes within an epoch (top) and the duration of each epoch (bottom) reveals a signature (dark green) averaged across subjects (teal).

Finally, we ask whether we can predict the perceived merit of a discovery by measuring the node’s theoretical ability to influence a body of knowledge due to its location within the gappy topology. We operationalize perceived merit in two ways: (i) a network-based measure and (ii) the receipt of a Nobel prize (Szell et al., 2018; Li et al., 2019; Jin et al., 2021), although we acknowledge the imperfect nature of the latter assessment (The Lancet, 2018). After constructing a single network containing all nodes from all subjects, we calculated the network’s impulse response as a measure of a node’s potential influence. Arising from dynamical systems theory, the impulse response quantifies how much a network “responds” to an “impulse” that perturbs one node (Figure 5.4A; Section 5.7). We observed that nodes that more frequently participate in the birth or the death of cavities have higher impulse responses (birth: Pearson’s correlation coefficient  $r = 0.36$ ,  $p \ll 0.001$ ; death:  $r = 0.38$ ,  $p \ll 0.001$ ; Figure 5.4B). Importantly, such nodes are also more frequently awarded Nobel prizes (birth:  $KS = 0.15$ ,  $p = 4.2 \times 10^{-6}$ ; death:  $KS = 0.17$ ,  $p = 1.2 \times 10^{-7}$ ; Figure 5.4C-

D). Importantly, both operationalizations are higher-order measures, meaning that they do not depend only on the connections immediately around articles for Nobel prize-winning discoveries. These results do not imply causality in either direction yet suggest that the network structure may reflect the real-world influence of concepts on the existing body of knowledge.

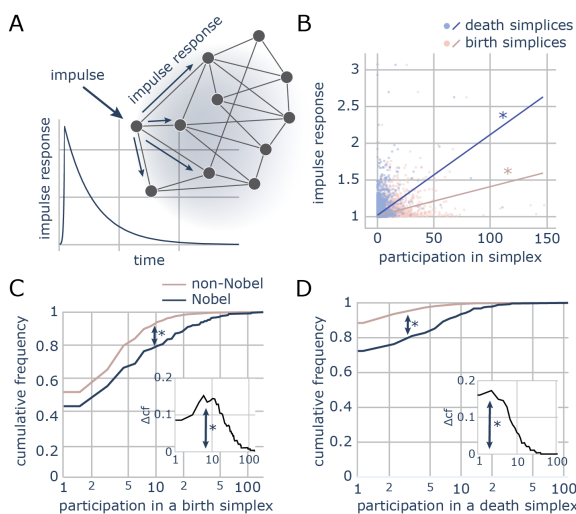


Figure 5.4: **Concept networks undergo a signature pattern in structural stability.** **A** Illustration of the response of a network to an impulse applied to a node. **B** Nodes that more frequently participate in the birth or death of persistent cavities have higher impulse response, which is a dynamical-systems measure of a node’s influence on a network (birth:  $r = 0.36$ ,  $p = 0$ ; death:  $r = 0.38$ ,  $p = 0$ ). **C-D** Nobel prizes are more frequently awarded for nodes that participate in the birth (panel C) and death (panel D) of persistent cavities (birth:  $KS = 0.15$ ,  $p = 4.2 \times 10^{-6}$ ; death:  $KS = 0.17$ ,  $p = 1.2 \times 10^{-7}$ ). Subpanels show the difference in cumulative frequencies ( $\Delta cf$ ).

In summary, our findings reveal that human knowledge, in the case of articles in Wikipedia, grows by filling gaps in knowledge, perhaps driven by the collective curiosity of individual scientists (Merton, 1974; Golman et al., 2018), through inward and outward exploration and gradual modifications to network structure. Moreover, knowledge discovered while creating and filling knowledge gaps is likely to be more influential and more frequently awarded in the scientific community.



## 5.5. Model Assessment

### 5.5.1. Feyerabend

While Feyerabend's theory arrived latest in chronological time (1975 vs. Kuhn's in 1962 and Lakatos's in 1970), we tested his hypothesis first because his ideas about the growth of scientific knowledge are the most different from the rest. Feyerabend posits that there is no single set of scientific methodologies employed in practice by all scientists (Feyerabend, 2010). Note that this claim stands in stark contrast to the Popperian and Kuhnian positions that science does have a characteristic pattern of discovery. Feyerabend was accepting of—and even promoted—a competition of theories and was careful to not demarcate science and its practice from other human endeavors, including storytelling and myth. These commitments served to counter the effects of a history of power structures and the influence of Western philosophy upon—and in justification of—science. To model Feyerabend's theory, we use an edge-rewiring process that produces a random network topology (Maslov et al., 2002). It is critical to note here that we are not modeling the scientific process as a process of random rewiring; instead, we are modeling the network structure resulting from science without a characteristic pattern of discovery (Feyerabend's thesis) as the culmination of a random rewiring of connections. In comparing edge-rewired networks to their real counterparts, we observed that real networks often display significantly non-random clustering on two topological scales: at the nodal scale, clustering manifests in a high clustering coefficient, whereas at the mesoscale, clustering manifests in the existence of modules and cores (Borgatti et al., 2000; Clauset, Newman, et al., 2004; Fagiolo, 2007). Thus, a network formulation of knowledge suggests that the processes underlying network growth constrain the network topology, in opposition to Feyerabend's theory.

Further, for Feyerabend, new significant discoveries in science come as a result of reframing or seeing nearly all things completely anew, rather than filling otherwise well-recognized gaps in knowledge. In contrast, in our study we have shown that novel discoveries are

the result of new cavities being formed or filled in the network and that these cavities are of higher-dimension, shorter period, and decreased frequency in real networks than in random or genetic networks. Hence, our results do not support the Feyerabend position of incommensurability and of novelty in knowledge being about a complete reconfiguration of the concept network. These observed distinctions between the data and Feyerabend's predictions motivate the examination of other philosophical positions.

### **5.5.2. Lakatos**

Lakatos hypothesized that science progresses as a research programme, which has a common “hard core” of postulates with an auxiliary belt of hypotheses that builds upon the core (Lakatos, 1968; Lakatos, 1978). He was interested in constructing a methodology of science (less so, an epistemology of science) that focuses less on demarcation (what counts as science or not) than on scientific practice. Within Lakatos's research programme, the “hard core” is a set of theories, practices, and commitments that most scientists would not want to give up in their research. Auxiliary hypotheses link the “hard core” to experiments and observations. Lakatos held that, in practice, science often chooses to modify auxiliary theories rather than give up on any of the “hard core” set of commitments, theories, and practices.

In this study, we operationalized Lakatos's hypothesis as growth within the core-periphery structure of a concept network (Borgatti et al., 2000). From the perspective of network topology, nodes in the “core” are densely connected to each other and form the topological center of the network; in contrast, nodes in the “periphery” tend to connect to the core but not to other peripheral nodes. In general, peripheral nodes have fewer connections than core nodes. The core-periphery structure of a network has important implications for the modification of scientific theories. For example, it is more difficult to modify nodes in the core than those in the periphery because core nodes are densely connected to one another. Similarly, it is easier to modify nodes in the periphery than those in the core because periphery nodes are only loosely connected to the network. The core-periphery

structure hence reflects Lakatos’s research programme with respect to both topology and scientific modification. In our study, we observed that concept networks do indeed display a core-periphery structure; however, we also observed that concept networks grow both “outward”, with core nodes preceding neighboring peripheral nodes, and “inward”, with core nodes preceded by neighboring peripheral nodes. This result supports an interesting aspect of Lakatos’s theory of science: it is seen as a layered core with an outer layer that is often updated by new discoveries that are influenced, in turn, by discoveries that occur in the periphery.

### 5.5.3. Kuhn

Kuhn’s ideas of scientific progress posit that there are two periods of science (Kuhn and Hacking, 2012). One period is called normal science in which scientists “solve puzzles” within the current view of the body of knowledge, which he called a paradigm. The other period is a paradigm shift in which the current paradigm is overturned by another. In our study, we first operationalized “puzzle-solving” normal science as filling knowledge gaps, and we formulated a conceptual gap as a topological cavity. Using persistent homology from the subfield of algebraic topology in mathematics, we identified cavities within a concept network and the times when cavities are created and destroyed throughout history (Zomorodian et al., 2005). We observed that cavities in real networks are filled more quickly than in edge-rewired networks or in genetic null models of knowledge growth. Moreover, real concept networks reflecting contemporary scientific knowledge have fewer unfilled cavities and more cavities with higher dimensions than edge-rewired or genetic null model networks. These results suggest that scientists create and fill knowledge gaps in the course of scientific progress.

To complement our casting of normal science as creating and filling cavities, we operationalized paradigms in terms of a concept network’s modular structure. Our choice was motivated by an appreciation of the following fact: the view that scientists hold about a body of knowledge can depend upon how a subject is organized into parts or modules. Con-

sider two examples. First, if a certain concept that originally exists in the fringe of a large module can engender enough discoveries, then the concept may start a new field of study and, at the same time, cause nodes to change their membership from the existing module to a new module. Second, when an incommensurable paradigm shift occurs, one might expect that knowledge is reorganized into new, unrecognizable modules. In operationalizing paradigms in terms of a concept network’s modular structure, we appreciate that paradigm shifts would manifest as shifts in modular structure. To detect such shifts, we formulated the growing concept network as a multilayer network wherein each layer contains the concept network as it existed in a given year. When we detected module membership across time (Mucha et al., 2010) and identified regimes of high or low change in module membership (Killick, Fearnhead, et al., 2012), we observed that concept networks displayed a signature pattern: a short period of little to no change, then a long period of small but constant change, then a short burst of many changes, and finally, a long period of little to no change. Importantly, these dynamics do not result in completely new modules—as one might expect with the “incommensurability” of paradigm shifts—but rather represent gradual changes to the existing modular structure.

Interestingly, the process of cavity filling in real networks reveals the importance of curiosity as a catalyst for scientific progress. Kuhn recognized quite early that “normal science” would push scientists (and knowledge formation) into a constrained and conservative path—one that does not encourage true creativity and innovation (Kuhn, 2000). For Kuhn, the innovation occurred either when iconoclastic individuals were lucky enough to uncover something novel, or when the accumulated failures of a scientific research programme made the search for more creative solutions necessary and better rewarded. In our study, we provide no analysis of individuals, scientific commitments, or practices. Yet, our results on cavity filling and its propensity for being rewarded in the scientific community suggest not two disparate periods of constrained normal science and innovative paradigm shifts but rather a single process in which scientists are continually and collectively driven, perhaps by curiosity, to uncover novel information that “connects the dots” among existing pieces of knowledge.

## 5.6. Discussion

The question of precisely how individuals and groups “do science” has troubled scholars for millennia. In the past century, philosophers of science have proposed several distinct and well-defined processes whereby scientific knowledge might grow. Prominent theories include those by Kuhn, Lakatos, and Feyerabend, which posit different patterns of growth that are supported by historical evidence. However, progress in devising rigorous empirical assessments of these theories has been hampered by difficulties in gathering large amounts of historical data and in operationalizing the theories in a falsifiable way. In this study, we formulated the body of knowledge as a concept network (Hesse, 1974) and operationalized philosophical theories in terms of the network’s growing structure. We demonstrated that, in the case of Wikipedia, concept networks are non-randomly organized, being characterized by high modularity and notable clustering. The networks do not grow strictly outward, as one might have naively expected. Rather, they expand both outward and inward. The inward expansion manifests as a filling of network cavities, which produces concepts that are topologically influential and more often awarded Nobel prizes. Across almost all subjects or fields, the modular structure of concept networks morphs along a signature trajectory of stability and instability consistent with Kuhn’s notion of a paradigm shift. Broadly, our mathematical formulations of historical data pave the way to describe, understand, and even potentially guide scientific progress for individuals and funding agencies (Fortunato et al., 2018). Furthermore, our findings provide a data-driven approach to identifying novel contributions, especially those by underrepresented groups whose works are typically devalued yet are vital for vibrant scientific innovation (Reardon, 2013; Hofstra et al., 2020).

In the remainder of this section, we briefly discuss our models of Kuhn’s, Lakatos’s, and Feyerabend’s theories using a complex systems approach, and we expand upon the insights revealed by our findings.

### 5.6.1. Feyerabend

While Feyerabend's theory arrived latest in chronological time (1975 vs. Kuhn's in 1962 and Lakatos's in 1970), we tested his hypothesis first because his ideas about the growth of scientific knowledge are the most different from the rest. Feyerabend posits that there is no single set of scientific methodologies employed in practice by all scientists (Feyerabend, 2010). Note that this claim stands in stark contrast to the Popperian and Kuhnian positions that science does have a characteristic pattern of discovery. Feyerabend was accepting of—and even promoted—a competition of theories and was careful to not demarcate science and its practice from other human endeavors, including storytelling and myth. These commitments served to counter the effects of a history of power structures and the influence of Western philosophy upon—and in justification of—science. To model Feyerabend's theory, we use an edge-rewiring process that produces a random network topology (Maslov et al., 2002). It is critical to note here that we are not modeling the scientific process as a process of random rewiring; instead, we are modeling the network structure resulting from science without a characteristic pattern of discovery (Feyerabend's thesis) as the culmination of a random rewiring of connections. In comparing edge-rewired networks to their real counterparts, we observed that real networks often display significantly non-random clustering on two topological scales: at the nodal scale, clustering manifests in a high clustering coefficient, whereas at the mesoscale, clustering manifests in the existence of modules and cores (Borgatti et al., 2000; Clauset, Newman, et al., 2004; Fagiolo, 2007). Thus, a network formulation of knowledge suggests that the processes underlying network growth constrain the network topology, in opposition to Feyerabend's theory.

Further, for Feyerabend new significant discoveries in science come as a result of reframing or seeing nearly all things completely anew, rather than filling otherwise well-recognized gaps in knowledge. In contrast, in our study we have shown that novel discoveries are the result of new cavities being formed or filled in the network and that these cavities are of higher-dimension, shorter period, and decreased frequency in real networks than in

random or genetic networks. Hence, our results do not support the Feyerabend position of incommensurability and of novelty in knowledge being about a complete reconfiguration of the concept network. These observed distinctions between the data and Feyerabend’s predictions motivate the examination of other philosophical positions.

### **5.6.2. Lakatos**

Lakatos hypothesized that science progresses as a research programme, which has a common “hard core” of postulates with an auxiliary belt of hypotheses that builds upon the core (Lakatos, 1968; Lakatos, 1978). He was interested in constructing a methodology of science (less so, an epistemology of science) that focuses less on demarcation (what counts as science or not) than on scientific practice. Within Lakatos’s research programme, the “hard core” is a set of theories, practices, and commitments that most scientists would not want to give up in their research. Auxiliary hypotheses link the “hard core” to experiments and observations. Lakatos held that, in practice, science often chooses to modify auxiliary theories rather than give up on any of the “hard core” set of commitments, theories, and practices.

In this study, we operationalized Lakatos’s hypothesis as growth within the core-periphery structure of a concept network (Borgatti et al., 2000). From the perspective of network topology, nodes in the “core” are densely connected to each other and form the topological center of the network; in contrast, nodes in the “periphery” tend to connect to the core but not to other peripheral nodes. In general, peripheral nodes have fewer connections than core nodes. The core-periphery structure of a network has important implications for the modification of scientific theories. For example, it is more difficult to modify nodes in the core than those in the periphery because core nodes are densely connected to one another. Similarly, it is easier to modify nodes in the periphery than those in the core because periphery nodes are only loosely connected to the network. The core-periphery structure hence reflects Lakatos’s research programme with respect to both topology and scientific modification. In our study, we observed that concept networks do indeed display

a core-periphery structure; however, we also observed that concept networks grow both “outward”, with core nodes preceding neighboring peripheral nodes, and “inward”, with core nodes preceded by neighboring peripheral nodes. This result supports an interesting aspect of Lakatos’s theory of science: it is seen as a layered core with an outer layer that is often updated by new discoveries that are influenced, in turn, by discoveries that occur in the periphery.

### 5.6.3. Kuhn

Kuhn’s ideas of scientific progress posit that there are two periods of science (Kuhn and Hacking, 2012). One period is called normal science in which scientists “solve puzzles” within the current view of the body of knowledge, which he called a paradigm. The other period is a paradigm shift in which the current paradigm is overturned by another. In our study, we first operationalized “puzzle-solving” normal science as filling knowledge gaps, and we formulated a conceptual gap as a topological cavity. Using persistent homology from the subfield of algebraic topology in mathematics, we identified cavities within a concept network and the times when cavities are created and destroyed throughout history (Zomorodian et al., 2005). We observed that cavities in real networks are filled more quickly than in edge-rewired networks or in genetic null models of knowledge growth. Moreover, real concept networks reflecting contemporary scientific knowledge have fewer unfilled cavities and more cavities with higher dimensions than edge-rewired or genetic null model networks. These results suggest that scientists create and fill knowledge gaps in the course of scientific progress.

To complement our casting of normal science as creating and filling cavities, we operationalized paradigms in terms of a concept network’s modular structure. Our choice was motivated by an appreciation of the following fact: the view that scientists hold about a body of knowledge can depend upon how a subject is organized into parts or modules. Consider two examples. First, if a certain concept that originally exists in the fringe of a large module can engender enough discoveries, then the concept may start a new field of study



and, at the same time, cause nodes to change their membership from the existing module to a new module. Second, when an incommensurable paradigm shift occurs, one might expect that knowledge is reorganized into new, unrecognizable modules. In operationalizing paradigms in terms of a concept network’s modular structure, we appreciate that paradigm shifts would manifest as shifts in modular structure. To detect such shifts, we formulated the growing concept network as a multilayer network wherein each layer contains the concept network as it existed in a given year. When we detected module membership across time (Mucha et al., 2010) and identified regimes of high or low change in module membership (Killick and Eckley, 2014), we observed that concept networks displayed a signature pattern: a short period of little to no change, then a long period of small but constant change, then a short burst of many changes, and finally, a long period of little to no change. Importantly, these dynamics do not result in completely new modules—as one might expect with the “incommensurability” of paradigm shifts—but rather represent gradual changes to the existing modular structure.

Interestingly, the process of cavity filling in real networks reveals the importance of curiosity as a catalyst for scientific progress. Kuhn recognized quite early that “normal science” would push scientists (and knowledge formation) into a constrained and conservative path—one that does not encourage true creativity and innovation (Kuhn, 2000). For Kuhn, the innovation occurred either when iconoclastic individuals were lucky enough to uncover something novel, or when the accumulated failures of a scientific research programme made the search for more creative solutions necessary and better rewarded. In our study, we provide no analysis of individuals, scientific commitments, or practices. Yet, our results on cavity filling and its propensity for being rewarded in the scientific community suggest not two disparate periods of constrained normal science and innovative paradigm shifts but rather a single process in which scientists are continually and collectively driven, perhaps by curiosity, to uncover novel information that “connects the dots” among existing pieces of knowledge.

#### 5.6.4. Network science

Our effort to quantitatively evaluate philosophical theories of science was made possible by formalizing large-scale data into well-defined expressions. In our formalization, we primarily employed network science and algebraic topology. Using the former, we extracted graphs from a large database of text in Wikipedia pages, and using the latter, we extracted simplicial complexes from the same. Network representations are intuitive models for concepts and conceptual relations. Such representations have previously proven useful in the study of concept networks from Wikipedia, and in the characterization of their topology using measures of centrality, shortest paths, and clustering (Bellomi et al., 2005; Matas et al., 2017; Lydon-Staley et al., 2021). Further, network representations can reveal patterns in data that are not quantifiable by observing each individual pairwise interaction—in this case, individual pages or hyperlinks—but that are only quantifiable by considering the entire network or subsection of a network. By complementing network science with algebraic topology, one can study higher order structures in concept networks, and thereby quantify the birth and death of topological cavities. This approach has previously proven useful in, for example, understanding the exposition of concepts in college textbooks (Christianson et al., 2020) and the growth of knowledge gaps in the semantic networks of toddlers (Sizemore et al., 2018)). By modeling concept networks as units and pairwise or even higher-order relationships among units, one can operationalize hypotheses about the structure of knowledge and its change over time.

Methodological assumptions and limitations. In operationalizing the complex social process of knowledge discovery, we made a several assumptions. Each was chosen to better enable us to form empirically testable hypotheses. These methodological assumptions and limitations are explained more extensively in Section 5.8 and are summarized here. First, to model knowledge discovery, we assume that the growing Wikipedia networks model how minds have collectively built networks of knowledge. Our study thus is not one of realism but one of the practice of science and scientific discovery (Hesse, 1980). Correspondingly, we

assume that collective minds, as represented by a Wikipedia network, have been shaped by evolutionary forces to reflect the real as closely as possible.

Furthermore, we acknowledge the limitations and assumptions of modeling scientific discovery as growing concept networks. First, the process of scientific discovery involves a network of commitments and social practices in addition to the concepts themselves (Longino, 2002). Thus, the findings presented here would best be considered complementary to qualitative studies of theory and process. Our approach is useful because it allows us to operationalize and quantitatively test certain dynamics in the structure of concept networks on a large scale. Second, while the network structure is different from the process used to obtain that structure, we aim to infer and constrain the possible processes that produced the concept network that we see today in the structure of Wikipedia. Indeed, concepts can evolve over time in their name or contents or can even be disconnected from the scientific canon (Shwed et al., 2010; Foster et al., 2015).

We note the limitations in our use of Nobel prizes as a measure of impact. As a highly subjective and often biased measure of impact (Lunnemann et al., 2019), we wish to test in future studies the robustness of our results not only in other datasets for Nobel prizes (Li et al., 2019) but also in other prizes (Jin et al., 2021). Moreover, while the goal of prizes is often to award for impact among other things, the effects of Nobel prizes on scientific practice is complex. Nobel prizes do not affect the citation impact of a Nobel laureate (Farys et al., 2017), but they do produce more papers and entrants in the topic associated with the scientific prize (Jin et al., 2021). Thus, scientific awards influence and are influenced by the body of scientific knowledge.

Finally, we acknowledge limitations in the way we represent Wikipedia pages as a concept network. First, without proper context, the title of a Wikipedia article can be ambiguous, and it may be difficult to infer the concept to which it best relates. This limitation is mitigated by the hyperlinks that connect a Wikipedia page to other pages; yet, even here it is important to acknowledge that the existence of a hyperlink involves some aspect of

chance. Second, Wikipedia pages only reflect the current understanding of a concept and its relation to other concepts. Third, Wikipedia itself may be laden with the predispositions of the editors of Wikipedia regarding their philosophies of science. Moreover, their editing of Wikipedia may be affected by their political biases (Harvard Business School et al., 2018) or their styles of information seeking, which may influence which articles they choose to read and later edit (Lydon-Staley et al., 2021). Fourth, Wikipedia articles may actually affect scientific practice itself, an observation that highlights the impact of secondary source information, especially a widely available source like Wikipedia, on primary research (Thompson et al., 2017). Thus, understanding patterns of growth on Wikipedia as they relate to primary sources seems ever more important. In the process of scientific development over the course of history, concepts themselves may change. In summary, we applied the network methods presented here to Wikipedia as an initial and interesting test case, but we look forward to future efforts applying such methods to other datasets, primary or otherwise.

#### **5.6.5. Future Directions**

In summary, we operationalize and test the theories of prominent philosophers of science using concept networks of Wikipedia articles. We significantly extend the state of the field by building upon previous studies that have examined the topology of Wikipedia (Bellomi et al., 2005; Matas et al., 2017; Yamada et al., 2020) and network principles in scientific practices (Sinatra et al., 2016; Clauset, Larremore, et al., 2017; Fortunato et al., 2018; Fontana et al., 2020). Moreover, we depart from prior work by using algebraic topology to test hypotheses regarding the sociology, history, and philosophy of science in a large-scale, empirical study. These methods can readily be applied to related questions about the structure and dynamics of concept networks. For example, differences in language or culture may influence the structure of concepts as encoded in Wikipedia pages. Other investigators may use network structure and document embedding to form more accurate and nuanced models of scientific discovery that explain the qualitative observations and interpretations of philosophers of science. Further, the methods demonstrated here may

reveal insights into observed gender disparities in the processes of knowledge generation and scientific engagement (Ford et al., 2017). Our development of data-driven approaches hence lays the groundwork for others to further describe and understand the process of scientific discovery and its relation to diversity, equity, and inclusion.

## **5.7. Methods**

Building growing concept networks from Wikipedia Software package for representing networks. We used the Python software package `networkx` (version 2.5) for most of our network representation and analysis. We used the Python package `igraph` (version 0.8.2) for representing networks for temporal module detection (see the later Section 5.7.9).

### **5.7.1. Selecting articles for a subject**

To build a concept network for a subject, we must first select Wikipedia articles that belong in a subject. Doing so in a principled manner can be difficult because there are no inherent delineations between articles of different subjects. Fortunately, Wikipedia provides indices of subjects, which list articles of a particular subject (<https://en.wikipedia.org/wiki/Wikipedia:Contents/Indices>). We chose to explore subjects in the areas of Mathematics and Logic, Natural and Physical Sciences, and Subdisciplines of Philosophy. For each subject, we built a network where nodes represented articles that are listed on the subject’s index, and where edges represented the articles’ hyperlinked connections.

### **5.7.2. Connecting articles (network nodes) via hyperlinks**

After gathering a list of articles to include in a subject-specific network, we first create a node for each article. Then to connect the nodes, we select hyperlinks that are in the lead section of an article (which we will call article A for illustration; [https://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style/Lead\\_section](https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section)). We take the hyperlinks from the lead section (i.e., the introduction) to (i) capture a concise overview of the concept, (ii) maintain

a normal distribution of the number of hyperlinks, (iii) reduce edge density, and (iv) avoid spurious hyperlinks to tangentially related pages. If the hyperlinks point to other articles in the subject, we then create a directed edge from those other articles to article A. We chose to direct edges from the hyperlinked article to the hyperlinking article because the hyperlinked article is used to explain the hyperlinking article and thus influences the information presented in the hyperlinking article. “Redirect” pages are redirected to the final redirected page, and internal links (links between Wikipedia pages) do not have disambiguation. After constructing the networks, we characterized their topological structure using network measures; in Table 5.1 and Table 5.2 we summarize the network measures that we used.

### 5.7.3. Weighting network edges

We determine the weight of each edge between two articles by calculating the cosine similarity between the vector of term frequency-inverse document frequencies (*tf-idf*) for words in one article and the *tf-idf* vector for words in the other article (Salton et al., 1975). We compute *tf-idf* by multiplying a local component (term frequency) with a global component (inverse document frequency). The measure is defined as follows:

$$\text{tfidf}_{i,j} = \text{frequency}_{i,j} \times \log_2 \frac{D}{\text{document\_frequency}_i},$$

where for term  $i$  in document  $j$ ,  $\text{frequency}_{i,j}$  is the number of times that term  $i$  occurs in document  $j$ ,  $D$  is the number of documents, and  $\text{document\_frequency}_i$  is the number of documents that contain term  $i$ . The *tf-idf* is a product of a token’s frequency and the token’s inverse document frequency. Thus, common tokens appearing very frequently in the corpus will be down-weighted whereas rare terms will be up-weighted. To account for differences in document length, we applied a common normalization such that the Euclidean norm of the *tf-idf* vector for a document became 1. After calculating the normalized *tf-idf* for each token, we quantified the similarity between pairs of nodes by computing the cosine

similarity between pairs of vectors. Thus, reciprocal links have the same weights. The cosine similarity results in a quantification of node similarity ranging from 0 to 1; higher values indicate greater similarity of the text between two Wikipedia pages. We use all articles in Wikipedia as the corpus for the calculation. We use the Python package `gensim` to compute *tf-idf*.

#### 5.7.4. Denoting the birth year of a node

We parse the years from the lead section and from a history section if the article has one. We denote the earliest year as the year when the node was “born” or conceived. There exist articles that do not have years listed in either the lead or the history section. To assign years to these articles, we first select all nodes without years whose parents (i.e., nodes with edges that link to a node) have years. For each such node, we denote its year as the year after the latest year of its parents. Then, we do the same for the remaining nodes without years. If a node still does not have a year, then we denote its year as 2020. The year of each edge is the latest year of either of its nodes. In a sensitivity analysis, we confirmed that our results were robust to small variations in the chosen year, suggesting that our findings were not unduly dependent on the specific algorithmic approach we employed (Figure 5.13). In a further validation analysis, we manually inspected the identified years. Specifically, we randomly sampled the passages containing the date from 40 pages in the following topics: biochemistry, cognitive science, evolutionary biology, genetics, molecular biology, energy, optics, philosophy of language, philosophy of law, philosophy of science, linguistics, software engineering.

#### 5.7.5. Parsing years

To better understand how we extracted the birth year of a node, here we provide additional detail about our algorithmic approach. Specifically, to parse the years from the text, we use regex to identify numbers that are preceded by months (e.g., “January”), prepositions of time (e.g., “around”), conjugations (i.e., “and”), articles (i.e., “the”), and other time-related

words (i.e., “early”, “mid”, “late”) and followed by the words BC, BCE, or MYA. We also parse centuries (e.g., “19th Century”) and convert them into years (e.g., 1800). We apply a negative sign to all years or centuries followed by BC or BCE for convenience in analysis, such that 1600 BC would become -1600. For Python implementation, see the function `filter_years(text)` in `module/wiki.py` in the code repository.

### 5.7.6. Null networks

To model Feyerabend’s hypothesis of anarchical scientific progress, we used an edge-rewiring process to construct null networks that could then be compared to real networks. Note first that we refer to the two nodes that an edge connects as the origin node and the target node. The rewiring process proceeds by first taking each edge and randomly selecting a new target node exactly once. The process is similar to that developed by Maslov and Sneppen (Maslov et al., 2002), but differs in that it does not swap the targets between pairs of nodes. In the context of Wikipedia articles, since we create an edge from a hyperlinked article to a hyperlinking article, we are maintaining the connection of an edge from the hyperlinked article but changing the connection of the edge to the hyperlinking article. Thus, we maintain degree distributions (Figure 5.9) while removing other features of the network topology.

### 5.7.7. Network methods

The network measures we used are summarized in Table S1. To calculate clustering coefficients, we convert a network into a weighted, undirected network and compute the clustering coefficients of each node in the network with the Python library `network` (Fagiolo, 2007). To calculate modularity, we also convert a network into a weighted, undirected network and maximize the following modularity quality index,

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(g_i, g_j),$$



where  $m$  is the number of edges in a network,  $A$  is the adjacency matrix of the network,  $i$  and  $j$  are nodes,  $g_i$  is the index of the community (or module) to which node  $i$  belongs, and  $\delta(g_i, g_j)$  is 1 if  $g_i$  and  $g_j$  are equal and 0 otherwise. We computed modularity using a greedy algorithm implemented by the Python library `networkx` (Clauset, Newman, et al., 2004). Finally, to calculate coreness, we convert a network into an unweighted, undirected network and use the `brain connectivity toolbox` (Rubinov and Sporns, 2010) to determine whether each node is a core or a periphery node, and to calculate the core-ness statistic of the network, which is given by

$$Q_C = \frac{1}{v_C} \sum_{i,j \in C_C} (w_{ij} - \gamma_C \bar{w}) - \sum_{i,j \in C_p} (w_{ij} - \gamma_C \bar{w})$$

where  $C_c$  is the set of all nodes in the core,  $C_p$  is the set of all nodes in the periphery,  $w_{ij}$  is the weight between nodes  $i$  and  $j$ ,  $\bar{w}$  the average edge weight,  $\gamma_C$  is a parameter that adjusts the size of the core, and  $v_C$  is a normalization constant (Rubinov, Ypma, et al., 2015). To compute the coreness of a thresholded network, we removed all edges with weights below the mean weight for a particular network and compute the coreness (**Fig. S11**).

### 5.7.8. Persistent homology

In our study, we hypothesized that processes of scientific discovery create and fill gaps in concept networks. A tool from applied algebraic topology, called persistent homology, provides a well-defined formulation of such gaps as persistent topological cavities that evolve as a network grows (Hatcher, 2002). To calculate persistent homology for a growing network, we first create a correspondence between  $k$ -cliques, which are all-to-all connected subnetworks of  $k$  nodes, and  $(k - 1)$ -simplices. Simplices of increasing dimension can be described as follows: a node is a 0-dimensional simplex, an edge is a 1-dimensional simplex, a 3-clique is a 2-dimensional simplex, and so on for higher dimensions. Using the Python package `dionysus2`, we add each clique as a simplex into a filtration at the latest year in the clique.

A filtration is formally a nested sequence of subspaces and can be thought of as a growing sequence of simplices. We may add more than one simplex to the filtration at each year, resulting in cavities with a lifetime of zero, which we remove for downstream analyses. In network terms, we are adding both nodes and edges across the filtration such that there may not necessarily be any disconnected components in a growing network. Finally, we use the package `dionysus2` to compute a reduced matrix which defines the indices of the start and end of cavities.

### 5.7.9. Temporally varying modularity

To detect modules across time, we first built a multilayer network from our growing networks. In the multilayer network, each layer is the network reflecting concepts discovered up to that year. Each node in a layer (e.g., at time  $t$ ) is connected to the equivalent node in the next layer (e.g., at time  $t + 1$ ) with weight 0.01. We empirically chose a low weight for the inter-layer links to ensure that we could effectively detect changes in modularity when nodes are only being added and not removed. We then used the Python implementation of Mucha et al. (2010) in the software package `leidenalg` (version 0.8.1) to compute the modules to which each node belongs. We set the parameter `partition_type` to `ModularityVertexPartition` to partition the network based on the optimization of a multilayer modularity quality function, with the `interslice_weight` set to 0.01, and the `n_iterations` set to -1; the latter choice directs the algorithm to run iterations until there is no improvement in the modularity quality index being optimized. The structural stability signature (Figure 5.3C) is robust to reasonable variations in the choice of the `interslice_weight` (i.e., keeping the weight on the order of  $10^{-2}$  or less) (Figure 5.10). Within this range of `interslice_weight` values, we find that only the magnitude—not the shape—of the signature changes.

In performing this analysis, we observed that the module membership of nodes changed at different rates across time, with almost no changes in module membership in the second half of the period evaluated. Hence, we hypothesized that there may be epochs of module

stability, with greater stability observed later in history. To identify such epochs, we first computed the number of changes in module membership across time as any time when the module membership is different than in the previous time point. Then, we summed the number of changes for each time point to obtain a single variable across time. Next, we used the R implementation of binary segmentation in the software package `changepoint` (Killick and Eckley, 2014). We used the function `cpt.meanvar` to detect changes in both mean and variance of the signal, which is the number of changes in module membership over time. We set the parameter `method` to “BinSeg” for binary segmentation, `Q` to 3 for the maximum number of changepoints, and `test.stat` to “Poisson” for a Poisson distribution. We chose these settings because the PELT segmentation algorithm, which selects the optimal number of changepoints `Q`, selected a `Q` of 3 for 14 out of 28 networks, with a median and a mode of 3 (Figure 5.11). Binary segmentation, on the other hand, allows the user to select a value for `Q`. So, we used binary segmentation with a `Q` of 3 for consistency across subjects. Additionally, we used a Poisson distribution because each change in module membership occurs independent of whether a node changes module membership in the previous time step. All other parameters were set to the default values. The algorithm then produces three indices, one for each changepoint, giving us four epochs (Figure 5.3B-C).

#### 5.7.10. Impulse response

To quantify the impact of an article-node on the subject network, we use the impulse response measure from linear systems theory (Kailath, 1980). Here, the network is represented as an adjacency matrix  $A$  such that an item  $a_{ij}$  in the matrix, with row  $i$  and column  $j$ , is the weight of the edge from the  $j^{th}$  node to the  $i^{th}$  node. Then, the impulse response of node  $i$  at time  $m$  is given by the  $i^{th}$  diagonal element of the controllability Gramian,

$$W_C = \sum_{m=0}^K A_{norm}^m B B^T (A_{norm}^T)^m,$$

where  $A_{norm}$  is normalized by dividing by one plus the dominant eigenvalue of  $A$ , and where  $B$  is a vector of ones (Kailath, 1980). Mathematically, this value quantifies the linearized response of the network to the activity of a node  $i$ . The activity of the node and the subsequent network response can be interpreted in the context of Wikipedia networks as a conceptual influence of node  $i$  on the rest of the network.

To quantify the influence of a node on all topics using the impulse response function, we built a large network that consisted of all nodes in all subjects under study. In addition, we took the impulse response to a time horizon  $m$  of 5 to capture up to five steps in the propagation of the impulse through the network. Because the dimensionality of the network is on the order of 105, it is also computationally prohibitive to compute  $A_m$  with larger values of  $m$ . Shorter time horizons also capture the relationship between impulse response and participation in the birth and death of cavities (Figure 5.12).

#### 5.7.11. Nobel prizes

We used Nobel prizes in Physics, Chemistry, and Physiology or Medicine as an external measure of influence. To identify which nodes in the concept networks received Nobel prizes, we parsed the Wikipedia articles “List of Nobel laureates in Physics”, “List of Nobel laureates in Chemistry”, and “List of Nobel laureates in Physiology or Medicine”. In the section “Laureates” for each article, there is a table of Nobel laureates that includes a rationale column, which describes the work of a laureate that motivated the Nobel prize with hyperlinks to articles that describe the laureate’s discoveries. For example, for the scientist Maria Skłodowska-Curie, the rationale column states, “for their joint researches on the radiation phenomena discovered by Professor Henri Becquerel” with a hyperlink to the article “Radiation”. By obtaining all hyperlinks to articles in the rationale columns, we identified nodes in the concept networks that were Nobel prize-winning nodes. All other nodes were identified as non-Nobel prize-winning.

#### **5.7.12. Data and code availability**

All code is available on <https://github.com/harangju/wikinet>. All data used in the study are publicly available on <https://dumps.wikimedia.org/enwiki>.

## REFERENCES

- Astegiano, Julia, Esther Sebastián-González, and Camila de Toledo Castanho (June 2019). “Unravelling the gender productivity gap in science: a meta-analytical review.” In: *Royal Society Open Science* 6.6, p. 181566. DOI: 10.1098/rsos.181566.
- Bellomi, Francesco and Roberto Bonato (2005). “Network analysis for Wikipedia.” In: *proceedings of Wikimania*, p. 81.
- Bianconi, Ginestra (2018). *Multilayer networks: structure and function*. First edition. Oxford, United Kingdom: Oxford University Press. 402 pp. ISBN: 978-0-19-875391-9.
- Borgatti, Stephen P and Martin G Everett (Oct. 2000). “Models of core/periphery structures.” In: *Social Networks* 21.4, pp. 375–395. DOI: 10.1016/S0378-8733(99)00019-2.
- Chemla, Karine and Evelyn Fox Keller, eds. (2017). *Cultures without culturalism: the making of scientific knowledge*. Durham: Duke University Press. 410 pp.
- Christianson, Nicolas H., Ann Sizemore Blevins, and Danielle S. Bassett (July 2020). “Architecture and evolution of semantic networks in mathematics texts.” In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 476.2239, p. 20190741. DOI: 10.1098/rspa.2019.0741.
- Clauset, Aaron, Daniel B. Larremore, and Roberta Sinatra (Feb. 3, 2017). “Data-driven predictions in the science of science.” In: *Science* 355.6324, pp. 477–480. DOI: 10.1126/science.aal4217.
- Clauset, Aaron, M. E. J. Newman, and Cristopher Moore (Dec. 6, 2004). “Finding community structure in very large networks.” In: *Physical Review E* 70.6, p. 066111. DOI: 10.1103/PhysRevE.70.066111.
- Daston, Lorraine (2016). “History of Science without Structure.” In: *Kuhn’s Structure of Scientific Revolutions at fifty: reflections on a science classic*. Ed. by Robert J. Richards and Lorraine Daston. Chicago: University of Chicago Press, pp. 115–132.
- Fagiolo, Giorgio (Aug. 16, 2007). “Clustering in complex directed networks.” In: *Physical Review E* 76.2, p. 026107. DOI: 10.1103/PhysRevE.76.026107.

- Farys, Rudolf and Tobias Wolbring (Sept. 2017). “Matched control groups for modeling events in citation data: An illustration of nobel prize effects in citation networks.” In: *Journal of the Association for Information Science and Technology* 68.9, pp. 2201–2210. DOI: 10.1002/asi.23802.
- Feyerabend, Paul (2010). *Against method*. 4th ed. London ; New York: Verso. 296 pp. ISBN: 978-1-84467-442-8.
- Fontana, Magda et al. (Sept. 2020). “New and atypical combinations: An assessment of novelty and interdisciplinarity.” In: *Research Policy* 49.7, p. 104063. DOI: 10.1016/j.respol.2020.104063.
- Ford, Heather and Judy Wajcman (Aug. 2017). “‘Anyone can edit’, not everyone does: Wikipedia’s infrastructure and the gender gap.” In: *Social Studies of Science* 47.4, pp. 511–527. DOI: 10.1177/0306312717692172.
- Fortunato, Santo et al. (Mar. 2, 2018). “Science of science.” In: *Science* 359.6379, eaao0185. DOI: 10.1126/science.aao0185.
- Foster, Jacob G., Andrey Rzhetsky, and James A. Evans (Oct. 2015). “Tradition and Innovation in Scientists’ Research Strategies.” In: *American Sociological Review* 80.5, pp. 875–908. DOI: 10.1177/0003122415601618.
- Golman, Russell and George Loewenstein (July 2018). “Information gaps: A theory of preferences regarding the presence and absence of information.” In: *Decision* 5.3, pp. 143–164. DOI: 10.1037/dec0000068.
- Harvard Business School et al. (Mar. 3, 2018). “Do Experts or Crowd-Based Models Produce More Bias? Evidence from Encyclopedia Britannica and Wikipedia.” In: *MIS Quarterly* 42.3, pp. 945–959. DOI: 10.25300/MISQ/2018/14084.
- Hatcher, Allen (2002). *Algebraic topology*. Cambridge ; New York: Cambridge University Press. 544 pp.
- Helmer, Markus et al. (Mar. 21, 2017). “Gender bias in scholarly peer review.” In: *eLife* 6, e21718. DOI: 10.7554/eLife.21718.

- Hesse, Mary B. (1974). *The structure of scientific inference*. London: Macmillan. 309 pp. ISBN: 978-0-333-15070-2.
- (1980). *Revolutions and reconstructions in the philosophy of science*. Harvester studies in philosophy 17,A. Brighton: Harvester Pr. 271 pp. ISBN: 978-0-85527-268-5.
- Hofstra, Bas et al. (Apr. 28, 2020). “The Diversity–Innovation Paradox in Science.” In: *Proceedings of the National Academy of Sciences* 117.17, pp. 9284–9291. DOI: 10.1073/pnas.1915378117.
- Jin, Ching, Yifang Ma, and Brian Uzzi (Dec. 2021). “Scientific prizes and the extraordinary growth of scientific topics.” In: *Nature Communications* 12.1, p. 5619. DOI: 10.1038/s41467-021-25712-2.
- Kailath, Thomas (1980). *Linear systems*. Prentice-Hall information and system science series. Englewood Cliffs, N.J: Prentice-Hall. ISBN: 978-0-13-536961-6.
- Killick, R., P. Fearnhead, and I. A. Eckley (Dec. 2012). “Optimal Detection of Changepoints With a Linear Computational Cost.” In: *Journal of the American Statistical Association* 107.500, pp. 1590–1598. DOI: 10.1080/01621459.2012.737745.
- Killick, Rebecca and Idris A. Eckley (2014). “**changepoint** : An *R* Package for Changepoint Analysis.” In: *Journal of Statistical Software* 58.3. DOI: 10.18637/jss.v058.i03.
- Kuhn, Thomas S. (2000). *The essential tension: selected studies in scientific tradition and change*. Chicago, Ill.: Univ. of Chicago Press. 366 pp. ISBN: 978-0-226-45806-9.
- Kuhn, Thomas S. and Ian Hacking (2012). *The structure of scientific revolutions*. Chicago ; London: The University of Chicago Press. 217 pp.
- Lakatos, Imre (1968). “Criticism and the Methodology of Scientific Research Programmes.” In: *Proceedings of the Aristotelian Society* 69, pp. 149–186.
- (1978). *The methodology of scientific research programmes*. Cambridge; New York: Cambridge University Press. ISBN: 978-0-511-62112-3.
- Li, Jichao et al. (Dec. 2019). “A dataset of publication records for Nobel laureates.” In: *Scientific Data* 6.1, p. 33. DOI: 10.1038/s41597-019-0033-6.



- Longino, Helen E. (2002). *The fate of knowledge*. Princeton, N.J: Princeton University Press. 233 pp.
- Lunnemann, Per, Mogens H. Jensen, and Liselotte Jauffred (Dec. 2019). “Gender bias in Nobel prizes.” In: *Palgrave Communications* 5.1, p. 46. DOI: 10.1057/s41599-019-0256-3.
- Lydon-Staley, David M. et al. (Mar. 2021). “Hunters, busybodies and the knowledge network building associated with deprivation curiosity.” In: *Nature Human Behaviour* 5.3, pp. 327–336. DOI: 10.1038/s41562-020-00985-7.
- Maslov, Sergei and Kim Sneppen (May 3, 2002). “Specificity and Stability in Topology of Protein Networks.” In: *Science* 296.5569, pp. 910–913. DOI: 10.1126/science.1065103.
- Matas, Neven, Sanda Martincic-Ipšić, and Ana Meštrović (Aug. 1, 2017). “Comparing Network Centrality Measures as Tools for Identifying Key Concepts in Complex Networks: A Case of Wikipedia.” In: *Journal of Digital Information Management* 15.4, p. 203. DOI: 10.6025/jdim/2017/15/4/203-213.
- Merton, Robert K. (1974). *The sociology of science: theoretical and empirical investigations*. 4. Dr. Chicago: Univ. of Chicago Pr. 605 pp. ISBN: 978-0-226-52092-6.
- Mucha, Peter J. et al. (May 14, 2010). “Community Structure in Time-Dependent, Multi-scale, and Multiplex Networks.” In: *Science* 328.5980, pp. 876–878. DOI: 10.1126/science.1184819.
- Nagaraj, Abhishek, Esther Shears, and Mathijs de Vaan (Sept. 22, 2020). “Improving data access democratizes and diversifies science.” In: *Proceedings of the National Academy of Sciences* 117.38, pp. 23490–23498. DOI: 10.1073/pnas.2001682117.
- Newton, Isaac (1675). *Isaac Newton letter to Robert Hooke*.
- Popper, Karl R. (2008). *The Logic of scientific discovery*. Repr. 2008 (twice). Routledge classics. London: Routledge. 513 pp. ISBN: 978-0-415-27844-7.
- Reardon, Jenny (2013). “On the Emergence of Science and Justice.” In: *Science, Technology, & Human Values* 38.2, pp. 176–200.
- Robinson-Garcia, Nicolas et al. (Oct. 28, 2020). “Task specialization across research careers.” In: *eLife* 9, e60586. DOI: 10.7554/eLife.60586.

- Rubinov, Mikail and Olaf Sporns (Sept. 2010). “Complex network measures of brain connectivity: Uses and interpretations.” In: *NeuroImage* 52.3, pp. 1059–1069. DOI: 10.1016/j.neuroimage.2009.10.003.
- Rubinov, Mikail, Rolf J. F. Ypma, et al. (Aug. 11, 2015). “Wiring cost and topological participation of the mouse brain connectome.” In: *Proceedings of the National Academy of Sciences* 112.32, pp. 10032–10037. DOI: 10.1073/pnas.1420315112.
- Salton, Gerard, Andrew Wong, and Chungshu Yang (Nov. 1, 1975). “A vector space model for automatic indexing.” In: *Communications of the ACM* 18.11, pp. 613–620. DOI: 10.1145/361219.361220.
- Shwed, Uri and Peter S. Bearman (Dec. 2010). “The Temporal Structure of Scientific Consensus Formation.” In: *American Sociological Review* 75.6, pp. 817–840. DOI: 10.1177/0003122410388488.
- Siew, Cynthia S. Q. et al. (June 17, 2019). “Cognitive Network Science: A Review of Research on Cognition through the Lens of Network Representations, Processes, and Dynamics.” In: *Complexity* 2019, pp. 1–24. DOI: 10.1155/2019/2108423.
- Sinatra, R. et al. (Nov. 4, 2016). “Quantifying the evolution of individual scientific impact.” In: *Science* 354.6312, aaf5239–aaf5239. DOI: 10.1126/science.aaf5239.
- Sizemore, Ann E. et al. (Sept. 2018). “Knowledge gaps in the early growth of semantic feature networks.” In: *Nature Human Behaviour* 2.9, pp. 682–692. DOI: 10.1038/s41562-018-0422-4.
- Szell, Michael, Yifang Ma, and Roberta Sinatra (Nov. 2018). “A Nobel opportunity for interdisciplinarity.” In: *Nature Physics* 14.11, pp. 1075–1078. DOI: 10.1038/s41567-018-0314-6.
- The Lancet (Oct. 2018). “The Nobel Foundation needs to check its privilege.” In: *The Lancet* 392.10154, p. 1168. DOI: 10.1016/S0140-6736(18)32359-6.
- Thompson, Neil and Douglas Hanley (2017). “Science Is Shaped by Wikipedia: Evidence from a Randomized Control Trial.” In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3039505.

- Wang, Dashun and Albert-László Barabási (2021). *The science of science*. Cambridge New York Port Melbourne New Delhi Singapore: Cambridge University Press. 303 pp. DOI: 10.1017/9781108610834.
- Yamada, Ikuya et al. (Sept. 26, 2020). “Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia.” In: *arXiv:1812.06280 [cs]*. arXiv: 1812.06280.
- Zeng, An et al. (Nov. 2017). “The science of science: From the perspective of complex systems.” In: *Physics Reports* 714-715, pp. 1–73. DOI: 10.1016/j.physrep.2017.10.001.
- Zomorodian, Afra and Gunnar Carlsson (Feb. 2005). “Computing Persistent Homology.” In: *Discrete & Computational Geometry* 33.2, pp. 249–274. DOI: 10.1007/s00454-004-1146-y.

# Supplementary Information

## 5.8. Supplementary Discussion

### 5.8.1. Assumptions and limitations of network models of scientific knowledge

Our primary tool for testing philosophical theories of scientific progress is the formalization of a body of knowledge into a network of concepts and their connections. This formalization allows the quantification of the structure of knowledge at different times in history and hence a data-intensive, statistical interpretation of philosophical theories (Leonelli, 2016). Here, we discuss the assumptions and limitations of network models of scientific knowledge.

### 5.8.2. Models of minds and of reality

In this study, we assume that the growing Wikipedia networks model how minds have collectively built networks of knowledge over the course of history and codified that knowledge in hyperlinked wikis. After Pierre Duhem’s criticisms of a simple Newtonian inductive method (i.e., the notion that science proceeds by generalizing from observations to theories) (Duhem et al., 1996) and Karl Popper’s “Critical Rationalism” and his method of “Falsification” to demarcate science from other forms of knowledge (Popper, 1968), most philosophers of science in the second half of the 20th century turned to descriptions of how Science (sometimes science without a capital “s” in Feyerabend’s case) works not only in theory but in practice. This turn sometimes included discovery and commitments to realism—the notion that there is a real world out there that science uncovers—but not always (Hesse, 1980). In various works, but in particular those of Kuhn and Lakatos, attention turned away from questions of realism and more towards the practice of science and of scientific discovery. Hence, the growth of knowledge is cast in this light and in the terminology of theoretical propositions, theoretical commitments, “hard core” theses, “auxiliary” hypotheses and experimental observations to describe changes in the practice of scientific research over historical time. Thus, reality sets the boundaries of what humans and their models can discover, but the two

spaces are not coterminous. Correspondingly, one can make the further assumption that the mind, or collective minds as represented by a Wikipedia network, has been shaped by evolutionary forces to reflect the real as closely as possible and lands upon one of several possible solutions in the space of the real.

### **5.8.3. Concepts versus commitments and practices**

A network of Wikipedia entries is different than a network of commitments, both theoretical and empirical, or of practices, which is often described by philosophers of science. Indeed, social and cultural values play an important role in the structuring of knowledge (Longino, 2002). Thus, we acknowledge that we formalize a body of knowledge only as its concepts and the interconnected relationships between the concepts without an account of theory, process, and observation. As such, the findings presented here would ideally be considered alongside qualitative studies of theory, process, and observation for a fuller picture of the structure of scientific knowledge. While Wikipedia networks themselves do not capture the discussions and theoretical commitments that are integral to scientific discovery, we add time as an additional dimension to our network analysis of concepts. Once we add time to networks, we can see changes in the structure of scientific knowledge over time, from which we may be able to quantitatively describe processes of scientific discovery.

### **5.8.4. Concepts as nodes**

In a concept network, each node represents a concept and is named after the title of a Wikipedia article. While Wikipedia articles in the hard sciences are relatively accurate (Giles, 2005), we acknowledge some limitations of representing concepts as the titles of Wikipedia articles. First, without the proper context, the title of a Wikipedia article by itself can be ambiguous with respect to the concept to which it refers. Second, not every Wikipedia article is a scientific concept; a Wikipedia article can be about other topics, such as scientists or scientific books. Both of these limitations in the representation of concepts, however, are mitigated by our network formulation. By linking nodes according

to their inter-dependence (see next paragraph), we provide a context for a concept-node with its neighboring concept-nodes. For example, the Wikipedia article on “On the Origin of Species” contains hyperlinks to “evolutionary biology”, “biodiversity”, “common descent”, “tree of life”, and “transmutation of species”, which help disambiguate and further define concepts in networks. Moreover, corpus-based semantic analysis can provide description and even explanation of semantics based on context (Berez et al., 2008). Lastly, while concepts themselves may change over the course of history, we use Wikipedia articles archived at a single time in their history, which runs this risk of obscuring how concepts slowly evolve to form new concepts (Figure 5.7B). As the work of conceptual history (Begriffsgeschichte) suggests, historical semantics—understanding where terms have come from and how their meanings have changed over time—is not only important for historical reasons, it also conditions contemporary thought and practice (Lloyd, 1994; Koselleck et al., 2002). Future study would ideally situate the concept-nodes examined here in cultural-linguistic context over time.

#### **5.8.5. Concept relationships as edges**

We formulate the relationship between concepts as the hyperlink between two Wikipedia articles. To illustrate the intuition for this formulation, suppose that there are two Wikipedia articles A and B. Article A hyperlinks to article B when article A uses in its description the word or concept that is described in article B. Thus, article B influences, or even in some case is necessary for, the description of article A, and we add an edge from article B to article A. Hence, a network of hyperlinked articles represents a network of influence or dependence between concepts. We acknowledge, however, that hyperlinks involve some aspect of chance, i.e., the willingness of an editor to write a linking content page and to hyperlink it. To ensure that only the most relevant hyperlinks are captured in the network models, we only use hyperlinks in the “lead” introductory section of each article.

### 5.8.6. Network structure versus process

The network structure is different from the process used to obtain that structure. In this study, we aim to infer the process that produced the concept network that we see today in the structure of Wikipedia. In fact, one may expect multiple processes to result in the same network structure. We therefore assume that a resultant network structure sets the boundaries to the types of processes that have built the network.

## 5.9. Supplementary Methods

### 5.9.1. Simulations of knowledge discovery

To examine the process of scientific discovery itself, rather than just the resultant structure of a concept network, we formulated a genetic model of knowledge discovery. A model of the discovery process is important for testing our hypothesis that the body of knowledge grows by creating and filling cavities. The intuition for the model is that it simulates scientists who learn about existing concepts and who slowly mutate them to form new concepts over the course of history. Importantly, the model has no “preference” (i.e., an objective or fitness function) that selects for certain features of new concepts. Thus, by comparing real networks to networks produced by the genetic model, we aim to explicitly test for whether real networks have a “preference” to fill knowledge gaps.

In simulations of knowledge discovery, we start with a subgraph of a subject network, consisting of nodes whose birth years are before 1 AD. For each subsequent year, we execute the following series of steps until either the year 2200 or the number of nodes of the model reaches that of the real subject-network: (i) initialize seeds for new nodes, (ii) mutate seeds, (iii) create new nodes for seeds, and (iv) connect new nodes to the model network. We capped the simulation at the year 2200 to halt the program in the case that the model takes a long time to, if not never, reach the number of nodes in the real subject-network.

### 5.9.2. Seed initialization

Before we begin to mutate a concept-node, we must first obtain a vector representation of a node to mutate. Thus, for all new nodes in the model network, we initialize a “seed”, which is a copy of the *tf-idf* vector of the “parent”, which is a node in the model network. As noted in the Methods section of the main text, the *tf-idf* vector of a node is the term-frequency inverse-document-frequency representation of the Wikipedia article corresponding to the node. Hence, by initializing and mutating the “seed”, we are slightly modifying a vector representation of a Wikipedia article with each iteration.

### 5.9.3. Seed mutation

With seed mutations, we model how scientist research a concept by taking a previously known concept and slightly modifying it to, for example, test a hypothesis about the known concept. Thus, to iteratively mutate a seed, we take three steps for each seed and for each year of the simulation: a point mutation, an insertion, and a deletion. Each step has a certain probability of occurrence, and we base those probabilities on the statistics that we find in real networks: a process that we explain in detail below.

First, a point mutation swaps the value of a randomly chosen element in the seed vector with a new value for each year with probability  $p$ . The new value is drawn from the distribution of *tf-idf* values in the original subject network (Figure 5.7). We set the probability  $p$  to approximate the change in *tf-idf* vectors over years in the real network. To approximate this change in *tf-idf* vectors, we first compute for each edge and its two nodes (i) the absolute difference in years, which we call the **year-diff**, (ii) the sum of the absolute difference between *tf-idf* values, which we call the **sum-abs-diff**, and (iii) the average absolute difference between 105 values randomly drawn from the original distribution of *tf-idf* values, which we call **avg-abs-diff**. Interestingly, there is a strong and statistically significant correlation between (i) **year-diff** and (ii) **sum-abs-diff** (Figure 5.7E). This correlation suggests that the longer the time between the discoveries of two neighboring nodes, the more different the



nodes are in their *tf-idf* representations. We thus approximate the probability of a point mutation  $p$  as the slope of the linear regression between (i) **year-diff** and (ii) **sum-abs-diff**, normalized by dividing by (iii) **avg-abs-diff**.

Second, an insertion randomly selects a zero element in the *tf-idf* vector and inserts a new value for each year with probability  $i$ . The new value is drawn from the distribution of *tf-idf* values in the real network. An insertion is thus equivalent to adding a word into an article. We set the probability  $i$  to approximate the change in words in an article over time. To calculate this probability, we first compute for each edge and its two nodes, (iv) the Manhattan distance between the two *tf-idf* vectors, which we call **man-dist** and which is intuitively the number of different words used between the two articles. Interestingly, there is a strong and statistically significant correlation between (i) **year-diff** and (iv) **man-dist** (Figure 5.7F). This correlation suggests that the longer the time between the discoveries of two neighboring nodes, the more different words are used in the two articles of each node. This relationship reflects the correlation between (i) **year-diff** and (ii) **sum-abs-diff** that was used to calculate the probability of point mutation  $p$  and points to the possibility of a slow-and-steady process underlying scientific research and discoveries.

Hence, we set the probability of insertion  $i$  as half of the slope of the linear regression between **man-dist** and **year-diff**. We use half of this slope for  $i$  because there is a third step for mutations that mirrors insertion: deletion. A deletion selects a randomly chosen non-zero element in the *tf-idf* vector and sets it to zero. Because a deletion of a *tf-idf* element is equivalent to removing one word in an article, just as insertion is equivalent to adding one word in the article, we set the probability of deletion  $d$  to the probability of insertion  $i$ .

#### 5.9.4. Node creation

In a real network, there is a distribution of cosine similarity values between *tf-idf* vectors of neighboring nodes; the mean of that distribution is around 0.3 (Figure 5.8B). In creating nodes from seeds in the simulation, we wished to match the distribution of cosine similarities

in the real network. To do so, when we initialize a seed, we draw a value from a normal distribution with a mean and standard deviation of the cosine similarities of the real network. Once the cosine similarity between a seed and its parent becomes less than the drawn value, we add the seed as a node in the network.

#### 5.9.5. Node connection

Once a node is added to a network, it must create connections to the rest of the network. To imitate the same process in a real Wikipedia article, we created a title of a new node, much like the title of a Wikipedia article. As the title, we selected ten words in each new node with the strongest *tf-idf* values excluding stop words. Then, if an existing node has a majority (i.e., six) of the words in the title, we create an edge from the new node to the existing node. The node will then be connected to the rest of the network, and in the next year of the simulation, the node will create a new seed on which to mutate.

#### 5.9.6. Statistical tests for core-periphery *lead-lag*

As a supplement to our analysis in Figure 5.1C, we performed two-sided one-sample t-tests for the *lead-lag* values for all core-periphery edges in each network with a null hypothesis that the mean is 0. Note that the *lead-lag* value is defined as the year of a core node minus the year of each of its neighboring peripheral nodes. Table 5.3 shows the t-statistic and p-values for t-tests for all subject-networks. A negative t-statistic indicates that core nodes are, on average, born before each of their neighboring peripheral node. Not all t-statistics are negative, and for subjects with negative t-statistics, the maximum p-value is 0.21. Thus, we cannot statistically conclude that in all subjects, core nodes on average precede their neighboring peripheral nodes.

### 5.9.7. Robustness of core-periphery *lead-lag*

To test for robustness of our core-periphery *lead-lag* analysis, we measured the core-periphery *lead-lag* (i) within modules, (ii) for all subjects, and (iii) across epochs. First, we tested whether knowledge indeed grows outward but within modules, in which case, the core-periphery *lead-lag* for an entire network would not reveal that core nodes lead peripheral nodes. By maximizing modularity (Newman et al., 2004) and performing core-periphery detection (Rombach et al., 2014), we can segregate nodes by modules as well as into cores and peripheries. Accordingly, we identified modules in each subject network using the Clauset-Newman-Moore greedy modularity maximization (Clauset et al., 2004). Within each module of each subject network, we identified the core and periphery nodes using the Borgatti-Everett core-periphery detection algorithm (Borgatti et al., 2000). Then, we computed the core-periphery *lead-lag* as the year of core nodes minus the year of neighboring peripheral nodes for each core-periphery edge. In accordance with our results in Fig. 1C of the main manuscript, we found that core nodes do not necessarily precede peripheral nodes within modules (Figure 5.5A).

Second, we computed the core-periphery *lead-lag* for all core-periphery edges in all networks to test whether core nodes precede peripheral nodes on average across all subjects (Figure 5.5B). We performed a two-sided t-test for the lead-lag values for all core-periphery edges in all networks with a null hypothesis that the mean is 0. We obtained a t-statistic of -58.9 ( $p \ll 0.001$ ). Thus, on average for nodes of all subjects, core nodes precede the periphery but clearly not always. This holistic statistical test supports a nuanced variation to Lakatos’s original hypothesis: that while the body of scientific knowledge does not have a “hard” core from which knowledge grows strictly outward, knowledge tends on average to grow outward from a “soft”, malleable core.

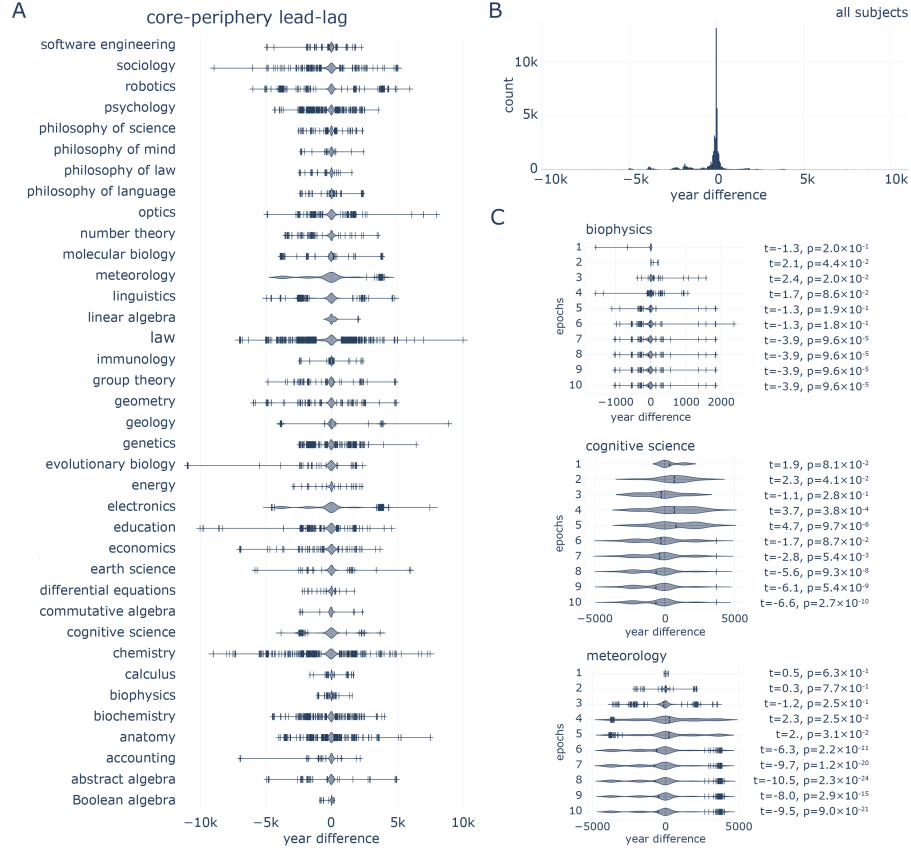
Third, we tested the core-periphery *lead-lag* in a time-dependent way to ensure robustness of our results to differences in core-periphery structure at earlier time points in a network’s

growth. Thus, for each subject network, we created subnetworks from the nodes that were present at ten time points in its history. The ten time points, or epochs, were equally spaced such that there is an equal number of unique years of nodes between each time point. As illustrated in three example subjects (Figure 5.5C), the t-statistic becomes more negative across epochs, suggesting that core nodes do not always precede peripheral nodes even with core-periphery structures of a network at earlier time points in its growth. Taken together, these supplementary analyses demonstrate the robustness of our core-periphery growth analysis within modules, for all subjects, and across epochs.

#### 5.9.8. Robustness of results to slight changes in years of nodes

Our analyses of cavity filling and the paradigm shift signature are sensitive to the discovery years of the nodes. Thus, to ensure the robustness of our results to slight changes in the estimated year of nodes, which could be due to any systematic or random errors, we again performed our analyses for the paradigm shift signature and persistent homology but now with networks that have years “jittered” by plus or minus one year. Assessing robustness is important especially given that not all Wikipedia articles have a history section or a year of discovery (Figure 5.13A). To jitter the years of real networks, we added a -1, 0, or 1, drawn uniformly with each number having a probability of 1/3, to the year of each node in all networks. We observed that the shape of the paradigm signature is robust to jittering in both the magnitude and duration (Figure 5.13B). In addition, we observed that our cavity-filling results are robust to jittering. The duration of cavities is slightly lower in jittered networks than in real networks ( $KS = 0.04$ ,  $p = 2.1 \times 10^{-4}$ ; Figure 5.13C). We note that this difference exists for short lifetimes on the order of 10 years, in contrast to the results shown in Figure 5.2D where the differences are present for longer lifetimes on the order of 100 and 1000 years. For the cavities that are currently present and for the dimensions of cavities, the cumulative frequencies are the same in the jittered networks as in the real networks (Figure 5.13D-E). These supplementary analyses demonstrate the robustness of our results on cavity-filling and paradigm shifts to small jittering of the years of nodes.

## 5.10. Supplementary Figures and Tables



**Figure 5.5: Core-periphery lead-lag relationship within modules, for all subjects, or across epochs.** **A** Core nodes do not necessarily precede their neighboring peripheral nodes within modules. Violin plots show distributions of the year of the core nodes minus the year of its neighboring peripheral nodes, for each core-periphery edge. Negative values indicate that core nodes are discovered before their neighboring peripheral nodes, and vice versa. Vertical lines indicate outliers, which are less than  $Q_1 - (1.5 \times IQR)$  or greater than  $Q_3 + (1.5 \times IQR)$  where  $Q_n$  is the  $n^{th}$  quartile and  $IQR = (Q_3 - Q_1)$ . **B** Distribution of core-periphery lead-lag for all core-periphery edges in all subjects. While core nodes do not always precede their neighboring peripheral nodes, core nodes do so on average ( $t = -58.9$ ,  $p \ll 0.001$ ; null hypothesis that sample mean is 0). **C** Core-periphery lead-lag plots computed for subnetworks at ten epochs in the history of each network for three example networks. Cores do not always precede their neighboring peripheral nodes even when core-periphery is calculated at previous time points in the history of a network.

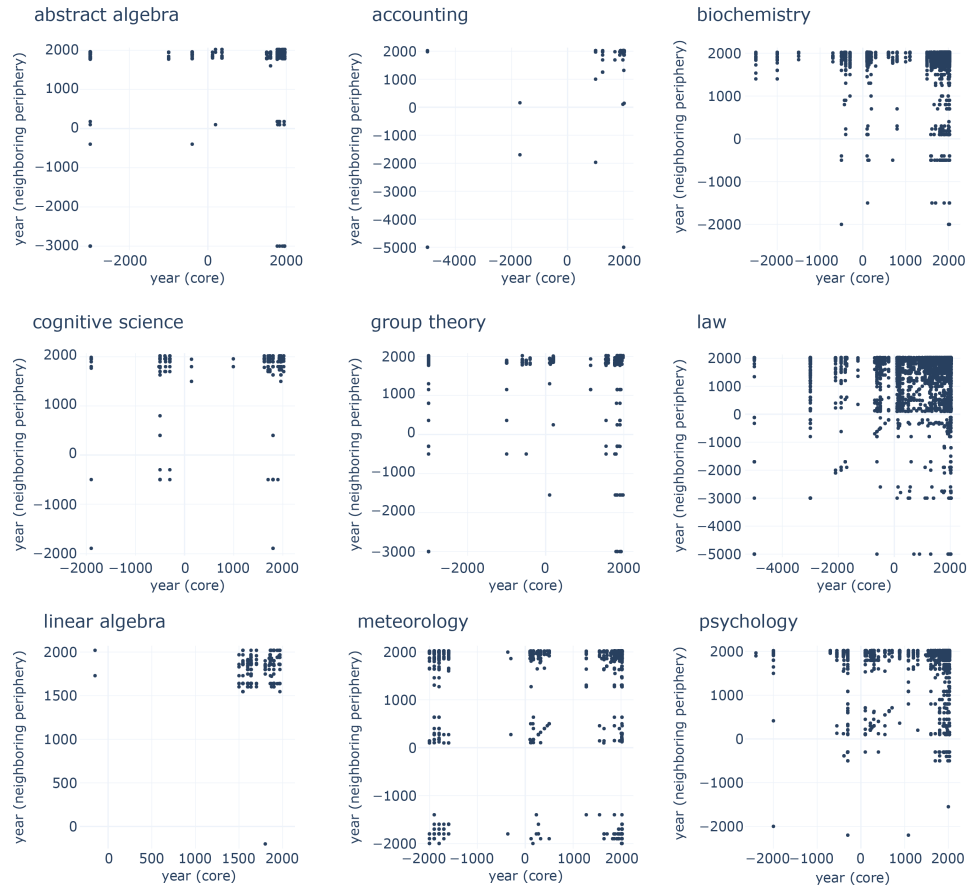


Figure 5.6: **Some core nodes are born early while other core nodes are born after their neighboring peripheral nodes.** While most core nodes both precede and follow their neighboring peripheral nodes, some core nodes are consistently born before neighboring peripheral nodes. The plots show the years for nodes in the core (x-axis) against the years for neighboring nodes in the periphery (y-axis) for nine example subjects. For each plot, a point on the top-left side indicates that a node in the core was born before a neighboring periphery node, whereas a point on the bottom-right side indicates that a node in the core was born after a neighboring periphery node. The peripheral nodes that are neighboring most core nodes are born both before and after their neighboring core nodes. Some core nodes (points on the very top left) are born before most, if not all, of their neighboring peripheries.

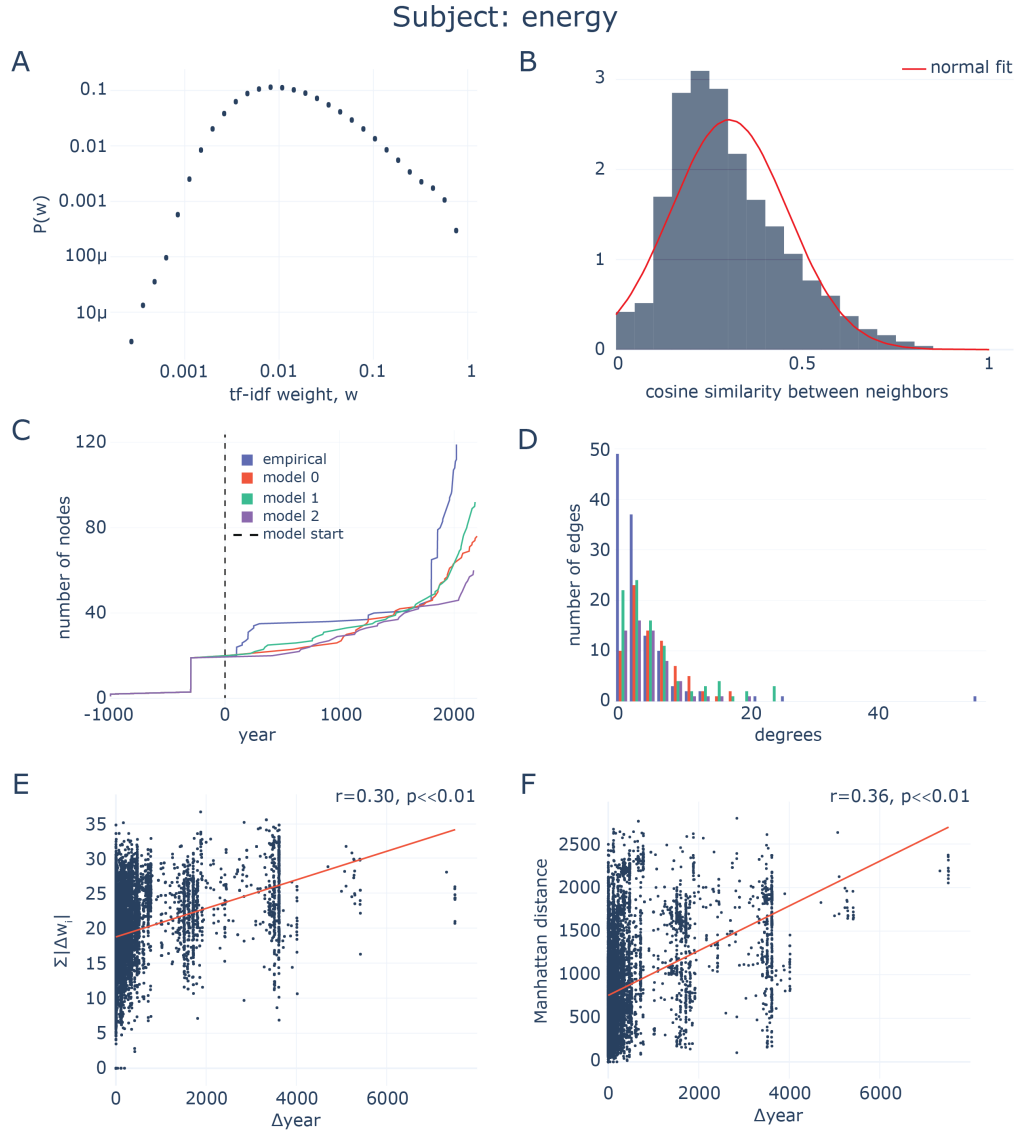


Figure 5.7: **Statistics of *tf-idf* vectors in an example real network.** For the example subject “Anatomy”, the statistics of the network used to inform simulations of knowledge discovery. **A** Distribution of *tf-idf* weights on a log-log plot. **B** Distribution of cosine similarity of *tf-idf* vectors between neighbors. **C** In three simulations, the model follows the growth of the real network and shows an exponential increase in the number of nodes. **D** The degree distributions in models are similar to that of the real network. **E** The sum of the absolute difference in *tf-idf* values between neighbors plotted against the year difference between neighbors. The correlation between the quantities suggests that more distant knowledge takes longer to discover. The vertical striations are due to temporal clusters of discovery; for example, in panel C, we can see that there are bouts of discoveries around the 3rd century BC, and around the 2nd and 17th centuries AD. Red line indicates line of best fit. **F** The Manhattan distance between *tf-idf* vectors between neighbors plotted against the year difference between neighbors.

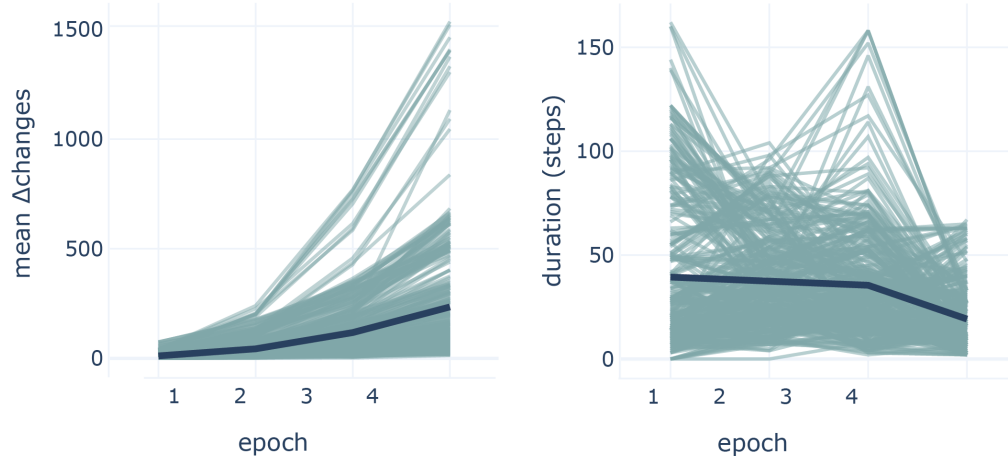


Figure 5.8: **Paradigm shift signature breaks in edge-rewired networks.** The mean number of changes within an epoch (left panel) and the duration of each epoch in time steps (right panel) reveals a different signature (dark green) in paradigm shifts across subjects (teal) in edge-rewired networks than that observed in the true data (see for comparison Figure 5.3C in the main text).

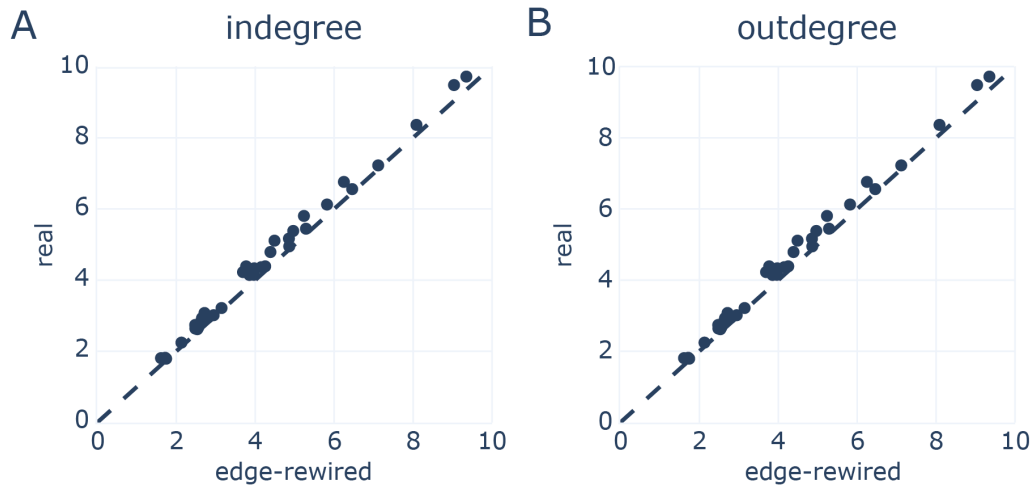
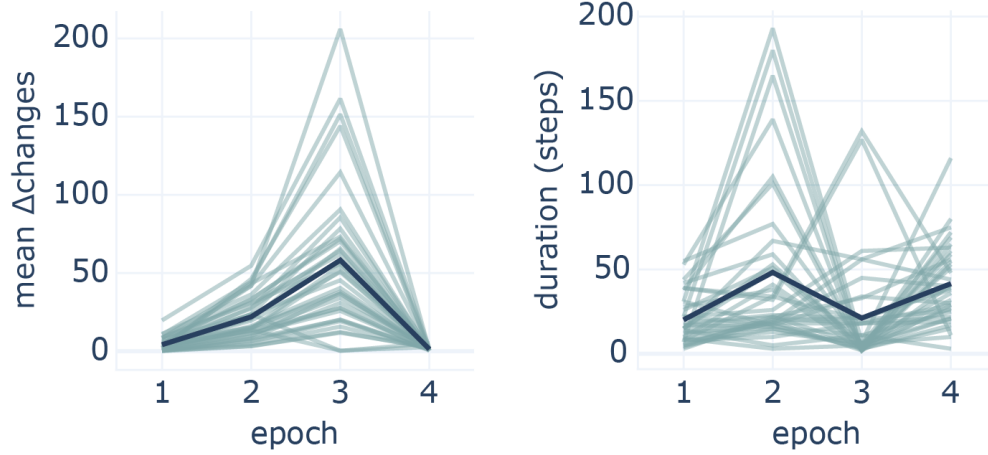


Figure 5.9: **Degree distributions of edge-rewired null networks.** A-B The average indegree (panel A) and average outdegree (panel B) of real concept networks compared to their edge-rewired nulls. Note that the average indegree and outdegree are identical because for any network, the total indegree is equivalent to the total outdegrees.



interslice weight = 0.001



interslice weight = 0.02

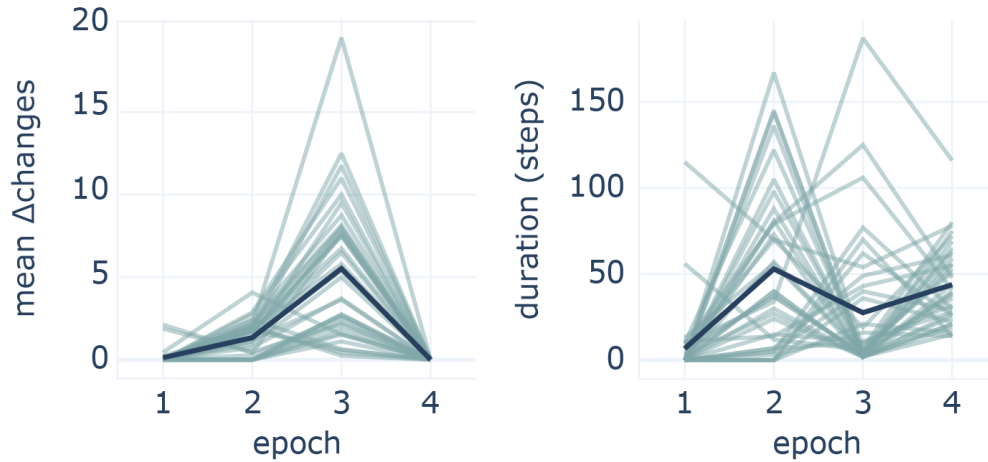
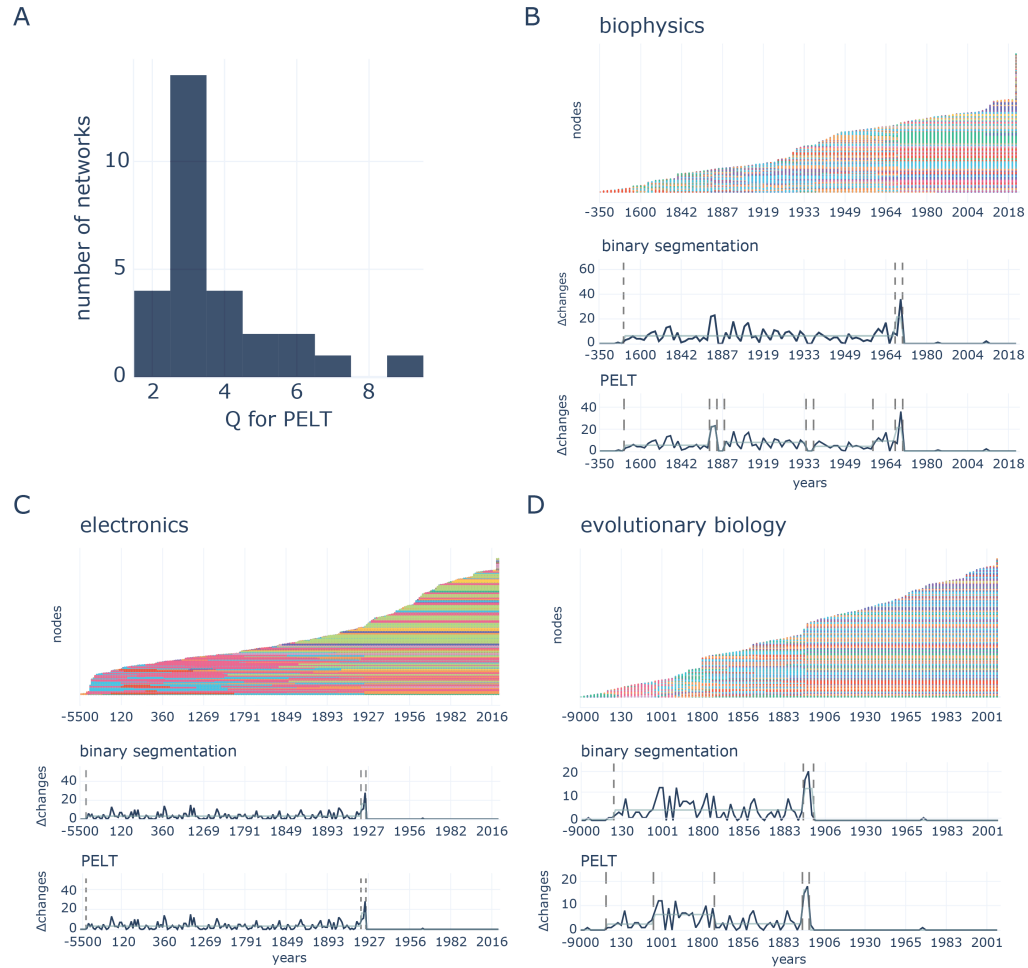


Figure 5.10: **The structural stability signature observed in Fig. 3 in the main text is robust to changes in the interslice weight.** The mean number of changes within an epoch (left panels) and the duration of each epoch (right panels) reveals a signature in paradigm shifts (dark green) averaged across subjects (teal) for interslice weights 0.001 (top panels) and 0.02 (bottom panels). For different interslice weights (top, bottom, and in Figure 5.3C in the main text), the magnitude—but not the shape—of the signature changes.



**Figure 5.11: Comparison of changepoint detection using PELT versus binary segmentation.** **A** Distribution of optimal number of changepoints  $Q$  discovered by the PELT algorithm for concept networks. The median and mode is 3. **B-D** Comparison of changepoint detection using PELT versus binary segmentation with a  $Q$  of 3 for the networks *biophysics* (panel B), *electronics* (panel C), and *evolutionary biology* (panel D). For each of the panels B, C, and D, the top plot shows the module membership across time points, and the bottom two plots show the number of changes in module membership. The grey dashed lines indicate changepoints detected by either binary segmentation or PELT, and the light green line indicates the average number of changes between each pair of changepoints. Most changepoints detected by binary segmentation with  $Q$  of 3 matches the changepoints detected by PELT with an optimal  $Q$ ; see, for example, the *electronics* network in panel C.

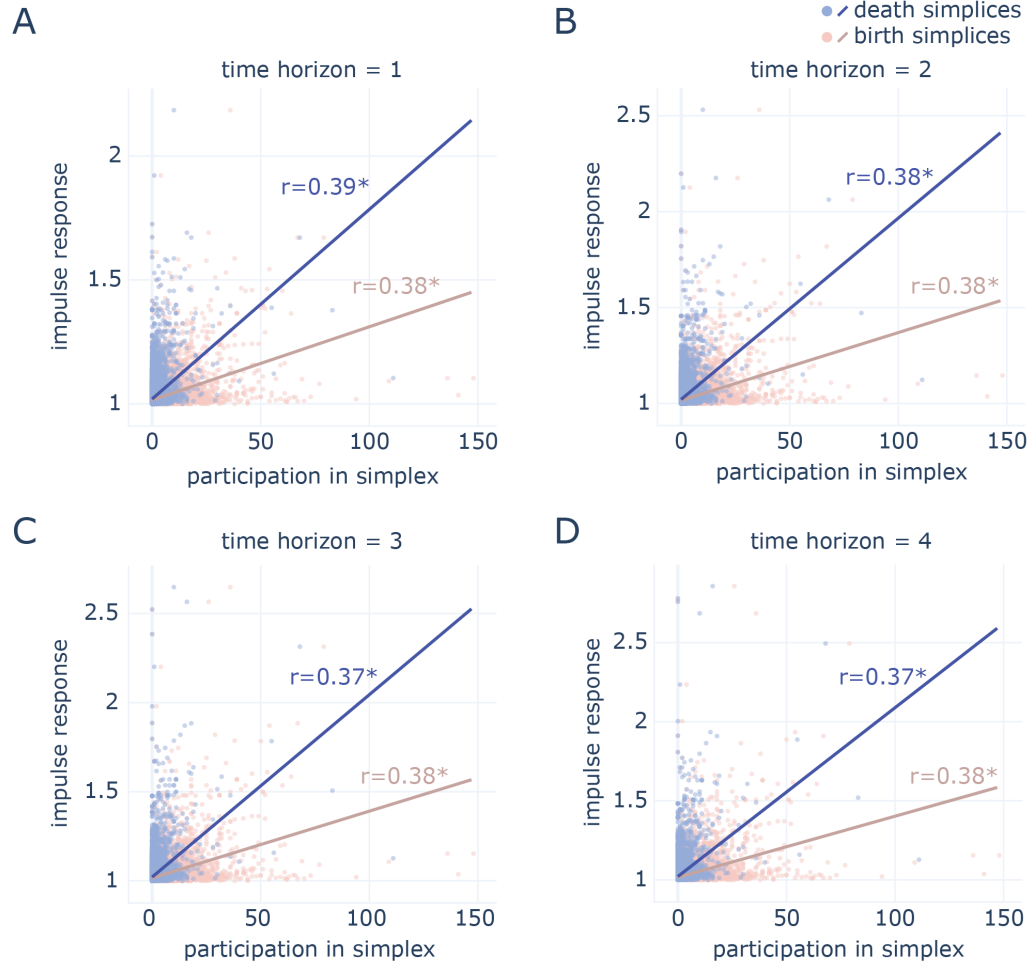


Figure 5.12: **The shorter time horizons capture the relationship between impulse response and participation in the birth and death of cavities.** A-D The correlation between impulse response of nodes and the frequency of participation of nodes in the birth and death of cavities is maintained across shorter time horizons (all  $p \ll 0.001$ ). A small decay in correlation, from around  $r = 0.38$  at a time horizon of 1 to  $r = 0.36$  at a time horizon of 5, demonstrates a robustness in the relationship between the cavity participation and the impulse response of nodes.

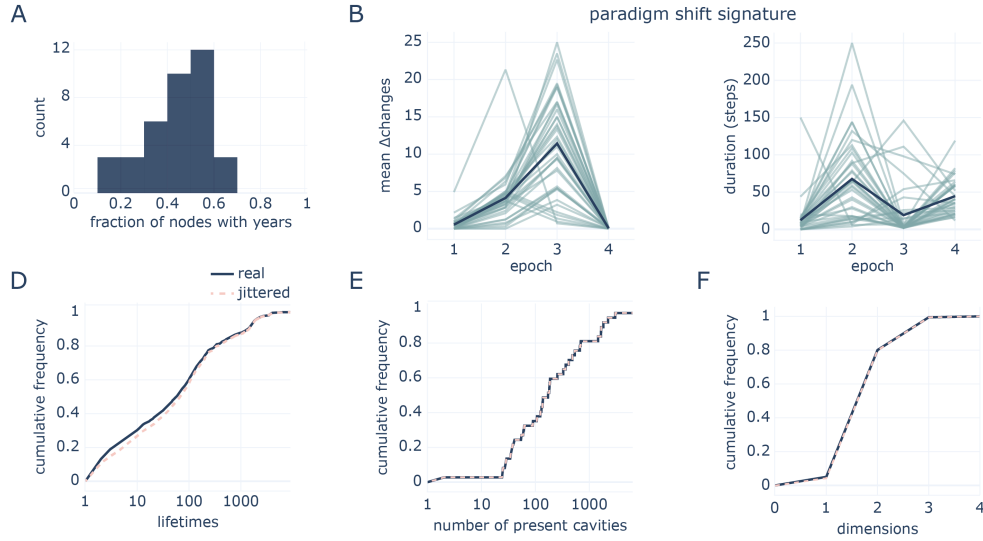


Figure 5.13: **Paradigm shift signature and cavity-filling statistics are robust to slight changes in the year of nodes.** **A** Distribution of the fraction of nodes whose articles have a year of discovery across subjects. **B** Signature in paradigm shifts (dark green) averaged across subjects (teal) in the mean number of changes within an epoch (left panel) and the duration of each epoch (right panel) for jittered networks. **C** Duration of knowledge gaps is slightly lower in jittered networks than in real networks ( $KS = 0.04$ ,  $p = 2.1 \times 10^{-4}$ ). We note that this difference exists for short lifetimes on the order of 10 years, which is in contrast to Figure 5.2D in the main text, where the differences are present for longer lifetimes on the order of 100 and 1000 years. **D** The frequency of knowledge gaps that are currently present (i.e., that have yet to die) is the same in jittered networks as in real networks. **E** The frequency of cavity dimensions is the same in jittered networks as in real networks.

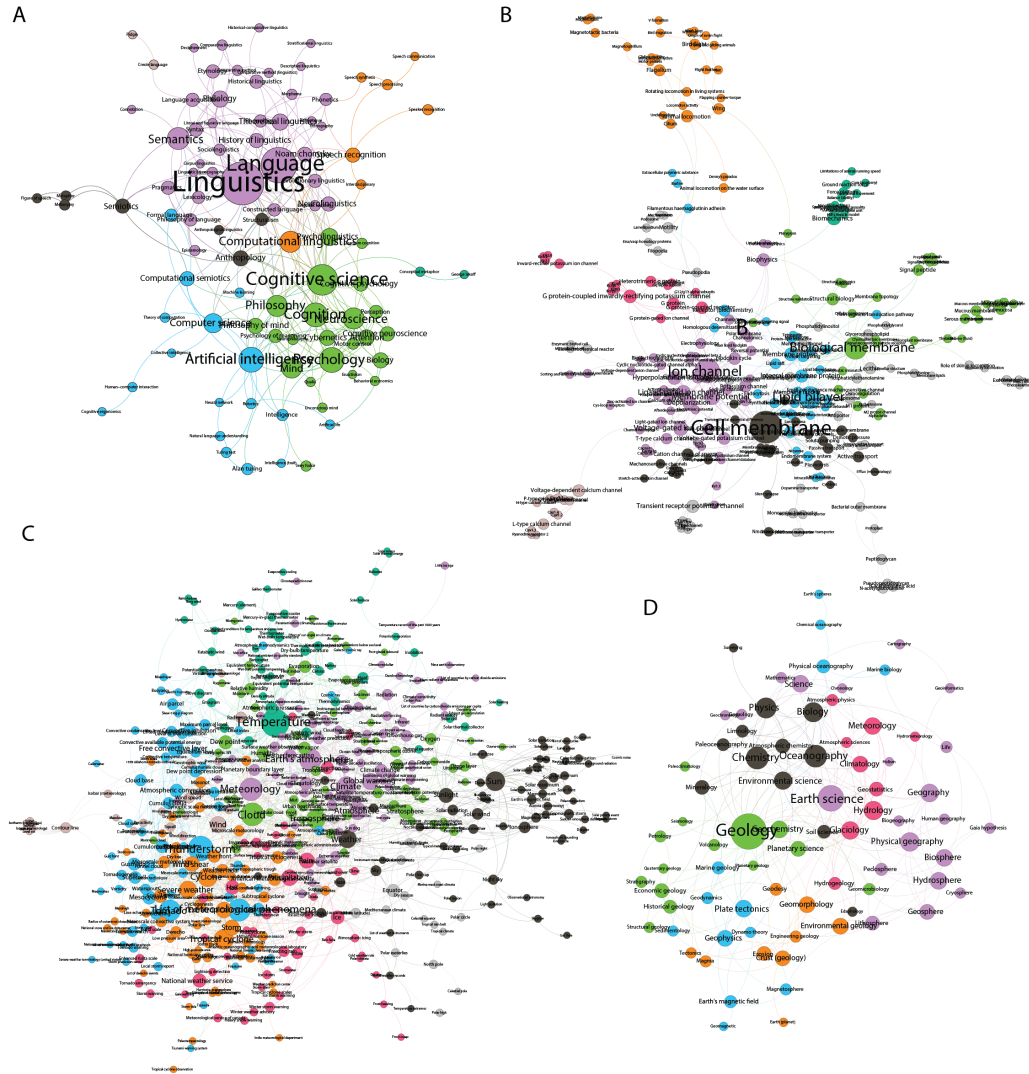


Figure 5.14: **Visualizations of four example networks.** **A** The network for the subject *cognitive science*. **(B)** The network for the subject *biochemistry*. **(C)** The network for the subject *metereology*. **(D)** The network for the subject *earth science*.

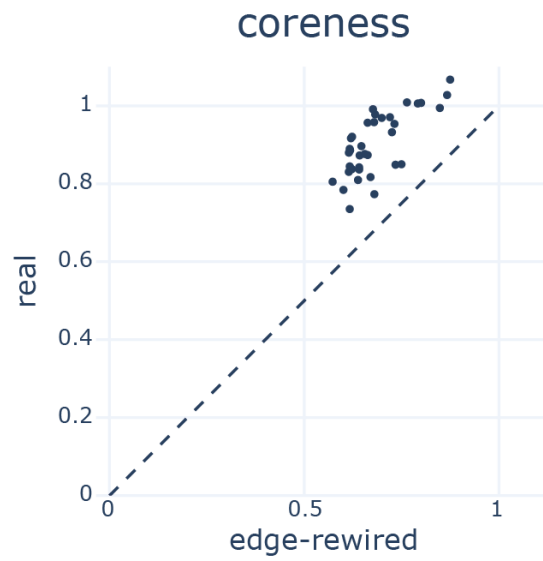


Figure 5.15: **Coreness of thresholded networks.** The final concept networks display greater coreness than edge-rewired null networks. The networks here are the same ones as in Figure 5.1C but are thresholded such that edges with weight below the mean weight for a particular network are removed.

Measure	Code	Equation
Clustering (Fagiolo, 2007)	<code>networkx.clustering()</code>	$c_n = \frac{2T(n)}{k_n(k_n-1)}$ <p>where <math>T(n)</math> is the number of triangles through node <math>n</math> and <math>k(n)</math> is the degree of node <math>n</math></p>
Modularity (Clauset et al., 2004)	<code>networkx.algorithms. community.modularity_max. greedy_modularity _communities()</code>	$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m} \delta(g_i, g_j)]$ <p>where <math>m</math> is the number of edges, <math>A</math> is the adjacency matrix, <math>i</math> and <math>j</math> are nodes, <math>g_i</math> is the index of the module to which node <math>i</math> belongs, and <math>\delta(g_i, g_j)</math> is 1 if <math>g_i = g_j</math> and 0 otherwise.</p>
Coreness (Rubinov et al., 2015)	<code>bct.core_periphery_dir()</code>	$Q_C = \frac{1}{v_C} \sum_{i,j \in C_C} (w_{ij} - \gamma_C \bar{w}) - \sum_{i,j \in C_p} (w_{ij} - \gamma_C \bar{w})$ <p>where <math>C_C</math> is the set of all nodes in the core, <math>C_p</math> is the set of all nodes in the periphery, <math>w_{ij}</math> is the weight between nodes <math>i</math> and <math>j</math>, <math>\bar{w}</math> is the average edge weight, <math>\gamma_C</math> is a parameter that adjusts the size of the core, and <math>v_C</math> is a normalization constant.</p>
Temporal modularity (Mucha et al., 2010)	<code>leidenalg.find _partition_temporal()</code>	$Q = \frac{1}{2\mu} \sum_{ijsr} [(A_{ijs} - \gamma_s \frac{k_{is} k_{js}}{2m_s}) \gamma_{sr} + \gamma_{ij} C_{jsr}] \delta(g_{is}, g_{jr})$ <p>where for nodes <math>i</math> and <math>j</math> in slices <math>s</math> and <math>r</math>, <math>2\mu = \sum_{jr} \kappa_{jr}</math>, <math>\kappa_{jr} = k_{js} c_{jr}</math>, <math>k_{js} = \sum_{is} A_{ijs}</math>, <math>c_{js} = \sum_r C_{jrs}</math>, <math>A_{ijs}</math> is the weight of the edge from <math>i</math> to <math>j</math> in <math>s</math>, <math>\gamma_s</math> is the resolution of slice <math>s</math>, <math>m_s = \sum_j k_{js}</math>, <math>\gamma(g_i, g_j)</math> is 1 if <math>g_i</math> and <math>g_j</math> are equal and 0 otherwise.</p>
Changepoint detection (Killick et al., 2014)	<code>cpt.meanvar()</code>	$ML(\tau_1) = \log p(y_{1:\tau_1}   \hat{\Theta}_1) + \log p(y_{(\tau_1+1):n}   \hat{\Theta}_2)$ <p>where <math>ML</math> is the maximum likelihood for a given change point at <math>\tau_1</math>, <math>y(1 : \tau_1)</math> is the signal from time 1 to <math>\tau_1</math>, and <math>\hat{\Theta}_1</math> is the maximum likelihood estimate of parameters (in our case the mean and variance).</p>

Table 5.1: Network measures.

Subject	N	Clustering	Modularity	Coreness
Boolean algebra	77	0.15±0.145	0.34	0.69
abstract algebra	356	0.18±0.116	0.36	0.73
accounting	115	0.14±0.158	0.44	0.68
anatomy	2043	0.12±0.104	0.56	0.72
biochemistry	1061	0.11±0.097	0.35	0.79
biophysics	463	0.10±0.161	0.59	0.94
calculus	103	0.19±0.153	0.38	0.67
chemistry	1032	0.13±0.100	0.26	0.82
cognitive science	115	0.22±0.148	0.25	0.72
commutative algebra	88	0.17±0.125	0.22	0.70
dynamical systems and differential equations	144	0.14±0.171	0.58	0.71
earth science	103	0.25±0.144	0.33	0.69
economics	511	0.11±0.118	0.42	0.72
education	668	0.10±0.148	0.53	0.91
electronics	1145	0.10±0.114	0.44	0.85
energy	119	0.11±0.156	0.44	0.83
evolutionary biology	265	0.16±0.118	0.26	0.78
genetics	1111	0.12±0.104	0.32	0.87
geology	92	0.14±0.144	0.31	0.75
geometry	294	0.17±0.129	0.38	0.75
group theory	295	0.18±0.129	0.29	0.76
immunology	410	0.14±0.164	0.45	0.89
law	3174	0.08±0.106	0.43	0.86
linear algebra	138	0.22±0.149	0.32	0.74
linguistics	395	0.16±0.119	0.35	0.76
meteorology	626	0.13±0.138	0.44	0.81
molecular biology	395	0.15±0.119	0.27	0.77
number theory	276	0.17±0.159	0.48	0.75
optics	352	0.16±0.135	0.31	0.79
philosophy of language	235	0.15±0.169	0.44	0.85
philosophy of law	146	0.09±0.149	0.47	0.81
philosophy of mind	106	0.19±0.161	0.28	0.75
philosophy of science	357	0.14±0.161	0.45	0.84
psychology	1568	0.09±0.123	0.45	0.86
robotics	1099	0.11±0.157	0.56	0.86
sociology	630	0.07±0.112	0.46	0.77
software engineering	226	0.14±0.132	0.36	0.71

Table 5.2: **Network metrics for subjects.**  $N$  is the number of nodes. Errors are standard deviations.



Subject	t-statistic	p-value
Boolean algebra	-1.73	$4.4 \times 10^{-2}$
abstract algebra	-11.70	$4.0 \times 10^{-30}$
accounting	-0.91	$1.8 \times 10^{-1}$
anatomy	-15.34	$1.7 \times 10^{-52}$
biochemistry	-15.30	$1.1 \times 10^{-51}$
biophysics	-3.94	$4.8 \times 10^{-5}$
calculus	-2.46	$7.4 \times 10^{-3}$
chemistry	-22.60	$7.3 \times 10^{-107}$
cognitive science	-6.62	$1.4 \times 10^{-10}$
commutative algebra	-3.03	$1.4 \times 10^{-3}$
dynamical systems and differential equations	-2.70	$3.9 \times 10^{-3}$
earth science	-0.82	$2.1 \times 10^{-1}$
economics	-4.94	$4.6 \times 10^{-7}$
education	-7.64	$2.8 \times 10^{-14}$
electronics	-19.83	$5.6 \times 10^{-82}$
energy	-2.40	$8.9 \times 10^{-3}$
evolutionary biology	2.64	$4.3 \times 10^{-3}$
genetics	-10.58	$3.9 \times 10^{-26}$
geology	-1.14	$1.3 \times 10^{-1}$
geometry	-10.81	$1.8 \times 10^{-25}$
group theory	-11.18	$3.8 \times 10^{-27}$
immunology	-6.16	$7.0 \times 10^{-10}$
law	-27.16	$2.9 \times 10^{-156}$
linear algebra	-1.87	$3.2 \times 10^{-2}$
linguistics	-13.01	$2.2 \times 10^{-36}$
meteorology	-9.44	$9.4 \times 10^{-21}$
molecular biology	-1.67	$4.7 \times 10^{-2}$
number theory	-8.17	$2.3 \times 10^{-15}$
optics	-8.26	$2.8 \times 10^{-16}$
philosophy of language	-2.94	$1.8 \times 10^{-3}$
philosophy of law	-3.75	$1.5 \times 10^{-4}$
philosophy of mind	-7.20	$1.1 \times 10^{-11}$
philosophy of science	-4.43	$5.7 \times 10^{-6}$
psychology	-16.58	$1.1 \times 10^{-59}$
robotics	-18.37	$6.8 \times 10^{-68}$
sociology	-5.93	$2.1 \times 10^{-9}$
software engineering	-3.53	$2.3 \times 10^{-4}$

Table 5.3: Core-periphery lead-lag t-tests for all subjects.

## REFERENCES

- Berez, Andrea and Stefan Gries (2008). “In defense of corpus-based methods: A behavioral profile analysis of polysemous get in English.” In: *Proceedings of the 24th Northwest Linguistics Conference*. DOI: 10.5167/UZH-84678.
- Borgatti, Stephen P and Martin G Everett (Oct. 2000). “Models of core/periphery structures.” In: *Social Networks* 21.4, pp. 375–395. DOI: 10.1016/S0378-8733(99)00019-2.
- Clauset, Aaron, M. E. J. Newman, and Cristopher Moore (Dec. 6, 2004). “Finding community structure in very large networks.” In: *Physical Review E* 70.6, p. 066111. DOI: 10.1103/PhysRevE.70.066111.
- Duhem, Pierre Maurice Marie, Roger Ariew, and Peter Barker (1996). *Essays in the history and philosophy of science*. Indianapolis: Hackett Pub. Co. 290 pp.
- Fagiolo, Giorgio (Aug. 16, 2007). “Clustering in complex directed networks.” In: *Physical Review E* 76.2, p. 026107. DOI: 10.1103/PhysRevE.76.026107.
- Giles, Jim (Dec. 2005). “Internet encyclopaedias go head to head.” In: *Nature* 438.7070, pp. 900–901. DOI: 10.1038/438900a.
- Hesse, Mary B. (1980). *Revolutions and reconstructions in the philosophy of science*. Harvester studies in philosophy 17,A. Brighton: Harvester Pr. 271 pp. ISBN: 978-0-85527-268-5.
- Killick, Rebecca and Idris A. Eckley (2014). “**changepoint** : An *R* Package for Changepoint Analysis.” In: *Journal of Statistical Software* 58.3. DOI: 10.18637/jss.v058.i03.
- Koselleck, Reinhart and Todd Samuel Presner (2002). *The practice of conceptual history: timing history, spacing concepts*. Cultural memory in the present. Stanford, Calif: Stanford University Press. 363 pp.
- Leonelli, Sabina (2016). *Data-centric biology: a philosophical study*. Chicago ; London: The University of Chicago Press. 275 pp.
- Lloyd, Elisabeth Anne (1994). *The structure and confirmation of evolutionary theory*. Princeton paperbacks. Princeton, N.J: Princeton University Press. 235 pp. ISBN: 978-0-691-00046-6.

- Longino, Helen E. (2002). *The fate of knowledge*. Princeton, N.J: Princeton University Press. 233 pp.
- Mucha, Peter J. et al. (May 14, 2010). “Community Structure in Time-Dependent, Multi-scale, and Multiplex Networks.” In: *Science* 328.5980, pp. 876–878. DOI: 10.1126/science.1184819.
- Newman, M. E. J. and M. Girvan (Feb. 26, 2004). “Finding and evaluating community structure in networks.” In: *Physical Review E* 69.2, p. 026113. DOI: 10.1103/PhysRevE.69.026113.
- Popper, Karl R. (1968). *Conjectures and refutations: the growth of scientific knowledge*. Harper torchbooks 1376. New York: Harper. 417 pp. ISBN: 978-0-06-131376-9.
- Rombach, M. Puck et al. (2014). “Core-Periphery Structure in Networks.” In: *SIAM Journal on Applied Mathematics* 74.1, pp. 167–190. DOI: 10.1137/120881683.
- Rubinov, Mikail et al. (Aug. 11, 2015). “Wiring cost and topological participation of the mouse brain connectome.” In: *Proceedings of the National Academy of Sciences* 112.32, pp. 10032–10037. DOI: 10.1073/pnas.1420315112.

## CHAPTER 6

### GENERAL CONCLUSIONS

#### 6.1. Results and overall discussion

In both brains and scientific knowledge, distributed representational systems represent information about the environment among distributed units and seek to form accurate models of the world on which to act. In biology, neural networks support the cognition that allows organisms, including humans, to identify and behave according to stimuli in the environment. In science, new concepts are discovered that build upon older concepts, thereby allowing humans to more accurately understand the physical and social world. Here, we first review the literature on network models and neural representations and then report the results of the three studies presented in this thesis. The first two studies contribute to our understanding of the processing of information performed by cascading neural networks, and the third study contributes to the field of science of science. In what follows, we will briefly summarize those contributions.

In the second chapter, we review dynamic representations in networked neural systems. We first survey network models of the brain, which seek to capture the interactions between neural units, from the scale of neurons themselves to voxels and brain regions. We then review a separate line of research in neuroscience that quantifies patterns in neural activity that correlate with environmental stimuli, such as faces (Adolphs, 2003). In our review, we propose that we may further our understanding of the neural correlates of cognition by uniting these two lines of research to study representations as they dynamically unfold on the underlying network. We discuss methods from other fields, such as algebraic topology and engineering, that may aid in our pursuit of these research goals.

In the third chapter, we study the relationship between the structure and dynamics of cascading neural networks. Cascading neural systems operate at a *critical* regime where

information is optimally stored and transmitted (Beggs and Timme, 2012; Wilting et al., 2019). It is unclear, however, whether and how these information processing properties are supported by network structure. We use methods from graph theory, linear systems theory, and information theory to explore the links between network structure, neural dynamics, and information processing. We find that neural networks achieve critical dynamics by tuning the system to have an eigenvalue around 1, thus propagating certain signals for long durations, and by using bi-directionally connected neurons, which have been widely observed in experiments.

In the fourth chapter, we broaden the scope of neural interactions from pairwise synaptic connections to triplet-wise interactions. Recent studies in information theory, called partial information decomposition, have begun to use triplet-wise network motifs to explore synergistic information that requires multiple inputs to determine the output (Wibral et al., 2017). While partial information decomposition measures synergistic information, it does not reveal the synergistic input-output mappings that it quantifies. In this study, we adapt the notion of logic gates from computer science to identify probabilistic logic gates in neural systems (Vahid, 2011). Neural logic gates reveal the synergistic computations that occur in neural systems and differ across brain regions and neural development.

In the fifth chapter, we explore the network science of historical concept networks. Philosophers of science have long theorized about the processes underlying scientific discovery. It has been challenging, however, to support or refute those theories with modern data analyses due to difficulties in systematizing large historical records. Here, we use Wikipedia, the largest online encyclopedia, and historical data in the articles to operationalize the growth of scientific knowledge as concept networks that grow throughout history. Using these methods, we find that knowledge does not grow outward but by filling in gaps in knowledge, which is more often influential and rewarded in the scientific community.

## 6.2. General limitations

In discussing the promise of the research presented here, some critical limitations should be noted. While neuronal avalanches have been widely observed (Beggs and Plenz, 2003; Beggs, 2004; Gireesh et al., 2008; Petermann et al., 2009; Hahn et al., 2010; Shriki et al., 2013; Bellay et al., 2015; Ponce-Alvarez et al., 2018; Shew et al., 2015), the methods that we have employed were specific to *in vivo* recordings of mouse and rat brains. As they differ across brain regions and across synaptic development, logic gates may differ *in vitro* recordings from those found *in vitro*, especially since neurons *in vivo* are receiving real stimulus from the environment rather than spontaneously firing. In our study, identifying neural logic gates require millisecond resolution of neural activity and tens to hundreds of neurons. Thus, to identify logic gates *in vivo*, any future studies with *in vivo* recordings must also have millisecond resolution with tens to hundreds of neurons.

Another general limitation of our analyses is the use of linear models in studying both neural and concept networks. In neural networks, while linear models are relatively good approximations of neural activity, neurons frequently exhibit non-linear responses to inputs (Nozari et al., 2021). Because neural firing is binary and stochastic, linear models are limited in that they can only approximate probabilities of firing at a resolution of milliseconds. In concept networks, we use linear models as an approximation of influence of one concept on another. This measurement assumes that one concept can influence another concept and that influence scales with textual similarity between documents describing the concepts. While the linear models of concept networks were complementary analyses of influence to our analyses using Nobel prizes, further study is required to characterize the influence from one concept to another.

In our science of science study, an important limitation was using Wikipedia as a database of science of history. While science articles in Wikipedia are similar in content to curated encyclopedias, such as Encyclopædia Britannica (Giles, 2005), biases remain in encyclopedic

articles (Ford et al., 2017; Harvard Business School et al., 2018). Interestingly, a recent study has found through a field experiment that Wikipedia influences scientific research itself (Thompson et al., 2017). Moreover, while the results of the study are robust to minor perturbations in the dates of discovery, the dates may vary based on the definitions of discovery; it is indeed a difficult problem, even for historians of science, to determine the exact date when a discovery was made. These limitations may be overcome by supplementing these analyses with publication data.

### 6.3. Future directions

In this section, we will discuss three complementary directions that our work can be taken in the future. A key feature of logic gates is that they are composable into machines that perform complex computations. In modern computer architectures, logic gates are used as blocks to build arithmetic logic units that ultimately make up the central processing unit (CPU) of a computer (Vahid, 2011). In neuroscience, neural correlates of cognition and behavior are studied separately from the physical properties of biological neural systems. Thus, it is critical to identify the building blocks of neural computation that can compose more complex computations performed by larger and larger population of neurons. Determining the atomic units of neural computation and the rules by which the units interact may prompt a new paradigm of studying neural systems.

A more immediate direction for studying neural logic gates is to expand logic gates from triplet-wise to even higher-order neural interactions. Recent studies have identified network motifs in neural networks, including hubs and rich-clubs, which coincidentally perform most of the computation in the local neuronal population (Markram, 1997; Song et al., 2005; Perin et al., 2011; Faber et al., 2019). Thus, it seems important to characterize higher-order neural interactions in which many neurons determine the activity of an individual neuron. We can characterize such interactions by using the information bottleneck method, through which we can determine which neurons are relevant to decode the firing of each neuron

(Murphy et al., 2022). Then, we can use probabilistic logic gates, which we developed in this dissertation, to identify the input-output mappings between neurons.

In our line of research in science of science, one future direction is to develop more accurate models of scientific discovery. In our fifth chapter, we develop a “genetic model” of scientific discovery in which we model scientists as randomly modifying concepts until they discover new concepts. In the science of science literature, Sinatra et al. (2016) demonstrates that the probability of a paper becoming highly successful, as measured by citations, is random across the publication history of individual scientists. A natural corollary may be that the probability of successful scientific discoveries itself is random, which would suggest that our “genetic model” is accurate in its description of the scientific process on the level of the individual but not on the level of a group of scientists. Thus, we can parameterize the model with group-level statistics—for example, the size of a group that is working on a topic—to control how quickly new concepts are discovered, which may be modeled with publication data or funding data. Future studies in this direction may further our understanding and control of the social and institutional processes underlying scientific discoveries.

## **6.4. Conclusions**

This work is unified by the application of network models and other computational tools to characterize two distributed representational systems: cascading neural networks and scientific progress. By applying these tools, we have gained insights into how networks govern the processing of information in their respective systems. In cascading neurons, we identified patterns in network structures, from the scale of two to hundreds of neurons, that support the dynamic storage and propagation of information. Moreover, we have begun to characterize how triplets of neurons interact to perform computation, which may, in the future, inform how computations can be composed into larger and more complex computations that support cognition. In science articles in Wikipedia, we used concept networks that grow across history to find that the body of knowledge grows by filling in



gaps in knowledge. In the future, these approaches may inform more accurate models of knowledge discovery for discoveries that have yet to be made.

## REFERENCES

- Adolphs, Ralph (Mar. 1, 2003). “Cognitive neuroscience of human social behaviour.” In: *Nature Reviews Neuroscience* 4, 165 EP.
- Beggs, J. M. (2004). “Neuronal Avalanches Are Diverse and Precise Activity Patterns That Are Stable for Many Hours in Cortical Slice Cultures.” In: *Journal of Neuroscience* 24.22, pp. 5216–5229. DOI: 10.1523/JNEUROSCI.0540-04.2004.
- Beggs, John M. and Dietmar Plenz (2003). “Neuronal Avalanches in Neocortical Circuits.” In: *Journal of Neuroscience* 23.35, pp. 11167–11177. DOI: 10.1523/JNEUROSCI.23-35-11167.2003.
- Beggs, John and Nicholas Timme (2012). “Being Critical of Criticality in the Brain.” In: *Frontiers in Physiology* 3, p. 163. DOI: 10.3389/fphys.2012.00163.
- Bellay, Timothy et al. (July 2015). “Irregular spiking of pyramidal neurons organizes as scale-invariant neuronal avalanches in the awake state.” In: *eLife* 4. Ed. by Frances K Skinner, e07224. DOI: 10.7554/eLife.07224.
- Faber, Samantha P. et al. (2019). “Computation is concentrated in rich clubs of local cortical networks.” In: *Network Neuroscience* 3.2, pp. 384–404. DOI: 10.1162/netn\_a\_00069.
- Ford, Heather and Judy Wajcman (Aug. 2017). “‘Anyone can edit’, not everyone does: Wikipedia’s infrastructure and the gender gap.” In: *Social Studies of Science* 47.4, pp. 511–527. DOI: 10.1177/0306312717692172.
- Giles, Jim (Dec. 2005). “Internet encyclopaedias go head to head.” In: *Nature* 438.7070, pp. 900–901. DOI: 10.1038/438900a.
- Gireesh, Elakkat D. and Dietmar Plenz (2008). “Neuronal avalanches organize as nested theta- and beta/gamma-oscillations during development of cortical layer 2/3.” In: *Proceedings of the National Academy of Sciences* 105.21, pp. 7576–7581. DOI: 10.1073/pnas.0800537105.
- Hahn, Gerald et al. (2010). “Neuronal Avalanches in Spontaneous Activity In Vivo.” In: *Journal of Neurophysiology* 104.6, pp. 3312–3322. DOI: 10.1152/jn.00953.2009.

- Harvard Business School et al. (Mar. 3, 2018). “Do Experts or Crowd-Based Models Produce More Bias? Evidence from Encyclopedia Britannica and Wikipedia.” In: *MIS Quarterly* 42.3, pp. 945–959. DOI: 10.25300/MISQ/2018/14084.
- Markram, H (1997). “A network of tufted layer 5 pyramidal neurons.” In: *Cerebral Cortex* 7.6, pp. 523–533. DOI: 10.1093/cercor/7.6.523.
- Murphy, Kieran A. and Dani Smith Bassett (2022). “The Distributed Information Bottleneck reveals the explanatory structure of complex systems.” In: *arXiv:2204.07576 [cs.LG]*. DOI: arXiv:2204.07576.
- Nozari, Erfan et al. (Aug. 11, 2021). *Is the brain macroscopically linear? A system identification of resting state dynamics*. arXiv: 2012.12351[cs, eess, math, q-bio].
- Perin, Rodrigo, Thomas K. Berger, and Henry Markram (2011). “A synaptic organizing principle for cortical neuronal groups.” In: *Proceedings of the National Academy of Sciences* 108.13, pp. 5419–5424. DOI: 10.1073/pnas.1016051108.
- Petermann, Thomas et al. (2009). “Spontaneous cortical activity in awake monkeys composed of neuronal avalanches.” In: *Proceedings of the National Academy of Sciences* 106.37, pp. 15921–15926. DOI: 10.1073/pnas.0904089106.
- Ponce-Alvarez, Adrián et al. (Nov. 2018). “Whole-Brain Neuronal Activity Displays Crackling Noise Dynamics.” In: *Neuron*. DOI: 10.1016/j.neuron.2018.10.045.
- Shew, Woodrow L. et al. (2015). “Adaptation to sensory input tunes visual cortex to criticality.” In: *Nature Physics* 11.8, pp. 659–663. DOI: 10.1038/nphys3370.
- Shriki, Oren et al. (2013). “Neuronal Avalanches in the Resting MEG of the Human Brain.” In: *Journal of Neuroscience* 33.16, pp. 7079–7090. DOI: 10.1523/JNEUROSCI.4286-12.2013.
- Sinatra, R. et al. (Nov. 4, 2016). “Quantifying the evolution of individual scientific impact.” In: *Science* 354.6312, aaf5239–aaf5239. DOI: 10.1126/science.aaf5239.
- Song, Sen et al. (Mar. 2005). “Highly Nonrandom Features of Synaptic Connectivity in Local Cortical Circuits.” In: *PLOS Biology* 3.3. DOI: 10.1371/journal.pbio.0030068.

- Thompson, Neil and Douglas Hanley (2017). “Science Is Shaped by Wikipedia: Evidence from a Randomized Control Trial.” In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3039505.
- Vahid, Frank (2011). *Digital design, with RTL design, VHDL, and Verilog*. 2nd ed. Hoboken, NJ: Wiley. ISBN: 978-0-470-53108-2.
- Wibral, Michael et al. (Mar. 2017). “Partial information decomposition as a unified approach to the specification of neural goal functions.” In: *Brain and Cognition* 112, pp. 25–38. DOI: 10.1016/j.bandc.2015.09.004.
- Wilting, J and V Priesemann (2019). “25 years of criticality in neuroscience — established results, open controversies, novel concepts.” In: *Current Opinion in Neurobiology* 58, pp. 105–111. DOI: 10.1016/j.conb.2019.08.002.