The why, how, and when of representations for complex systems

Leo Torres
leo@leotrs.com
Network Science Institute,
Northeastern University

Ann S. Blevins
annsize@seas.upenn.edu
Department of Bioengineering,
University of Pennsylvania

Danielle S. Bassett
dsb@seas.upenn.edu
Department of Bioengineering,
University of Pennsylvania

Tina Eliassi-Rad
tina@eliassi.org
Network Science Institute and
Khoury College of Computer Sciences,
Northeastern University

February 12, 2021

Contents

Abstract

Complex systems, composed at the most basic level of units and their interactions, describe phenomena in a wide variety of domains, from neuroscience to computer science and economics. The wide variety of applications has resulted in two key challenges: the progenation of many domain-specific strategies for complex system analyses that are seldom revisited or questioned, and the siloing of representation and analysis ideas within a domain due to inconsistency of complex systems language. In this work we offer basic, domain-agnostic language in order to advance towards a more cohesive vocabulary. We use this language to evaluate each step of the complex systems analysis pipeline, beginning with the system under study and data collected, then moving through different mathematical formalisms for encoding the observed data (i.e. graphs, simplicial complexes, and hypergraphs), and relevant computational methods for each formalism. At each step we consider different types of *dependencies*; these are properties of the system that describe how the existence of an interaction among a set of units in a system may affect the possibility of the existence of another relation. We discuss how dependencies may arise and how they may alter interpretation of results or the entirety of the analysis pipeline. We close with two real-world examples using co-authorship data and email communications data that illustrate how the system under study, the dependencies therein, the research question, and choice of mathematical representation influence the results. We hope this work can serve as an opportunity of reflection for experienced complex system scientists, as well as an introductory resource for new researchers.

1 Introduction

The term "complex system" is used to describe a multitude of systems of markedly different scales, from the atomic scale of interacting atoms to the vast scale of the whole universe, as well as markedly different behaviors, from starling murmurations to the viral spread of information on social media. Though distinct definitions exist, and not one is globally agreed upon, in general a complex system is (a) a collection of objects or agents with high cardinality, which (b) interact with one another in a non-trivial way, such that (c) the collective behavior of the system is unexpected, different than, or not immediately predictable from the aggregation of the behavior of the individual parts. This unique collective behavior is often said to *emerge* from the dynamics of the parts [103, 109]. For example, a population of neurons (units) connect via synapses (interactions) and consequently can perform computations (collective behavior). Additional real world examples include cellular reactions in photosynthesis, food webs in ecology, transactions in local markets, interconnected world-wide trading in economics, and various technologies such as the Internet and the power grid.

In order to study complex systems across disciplines and domains, it is important to concretely represent the system using a unifying mathematical language. In recent decades, the discipline of network science has arisen as the main focus of development of such a language [142]. Network scientists typically study complex systems by first modeling them using the tools and frameworks afforded by disciplines such as discrete mathematics and computational data structures. These formal frameworks, which we refer to as *formalisms* (see Section 1.1), enable the application of tried and true methodologies coming from different subfields within the mathematical, physical, and computational sciences. Furthermore, these formalisms allow for the execution of efficient algorithms and can be used to infer structure, function, and dynamics of a system. What makes this process somewhat challenging is that each encounter with a new complex system requires the construction of a new representation tailored to it. Network science is far from developing a single, unified representation that allows the study of all possible system structures and behaviors [115]. Indeed, there is currently not one, but a wealth of related frameworks, each of which captures particular perspectives and properties of the system under study.

This wealth of frameworks, and the resulting wealth of accompanying analysis pipelines, creates challenges for the study of complex systems. It hinders interdisciplinary communication, as researchers in one discipline may be unfamiliar with the representations and analyses used in another. Even within a single subfield, various approaches to represent and analyze the same complex system can hinder collective insight across research groups or projects [34]. As a consequence, it is difficult and sometimes impossible to gather insight across systems, which directly hampers the progress of complexity science [133]. As researchers striving for precision and efficiency, we must address this challenge by understanding the assumptions underlying each formalism, as well as the relationships between formalisms, and the impact of both formalism assumptions and relations on our analyses and interpretations of results.

In this work we aim to collect and align complex system analysis pipelines – from raw data procurement and clean-up to analysis results and final conclusions – while providing a common vocabulary for a continued discussion. While achieving a single, unified language is unlikely, we can at the very least begin to simplify and condense the pipelines currently in use. For clarity, we begin by defining the fundamental terms used throughout the paper. The main text follows the flow of Fig. 1, which illustrates a simplified representation of the analysis pipeline used when studying a complex system, insofar as it pertains to the formal representation of the system. We begin with an investigation of common system properties that can lead to biased analysis results if ignored, which we call *dependencies*, followed by definitions of three mathematical formalisms commonly used for representation. Next, we highlight mathematical relationships between formalisms that one might utilize in order to answer particular research questions, and finally we provide examples of computations suited for each of the three formalisms. Throughout the text we repeatedly ask how these dependencies and other modeling choices may influence the pipeline steps discussed. We provide two examples using a co-authorship dataset and the Enron emails dataset [23] to demonstrate the effects of various analysis pipelines on the results obtained from the same underlying system. Finally, we close by suggesting that each modeling decision in a research analysis pipeline be taken on a case-by-case basis and in consideration of the dependencies, formalisms, relationships, and research questions.

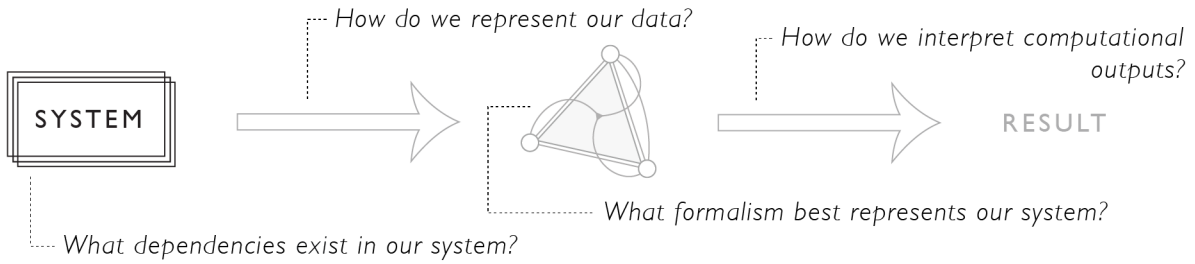

Figure 1: **Prototypical analysis pipeline for complex systems.** We begin with the system under study, and ask what sorts of elementary units exist, what relations exist that group elements together, and what dependencies might influence the existence of relations among units. We then turn to the question of how to represent the units, relations, and their dependencies; to answer this question, we must choose a formalism. Finally, we seek to interpret the outcomes of computations performed on the representation, and from those interpretations we reach a conclusion about the structure and function of the system.

1.1 Definitions

In this work we use a consistent language to allow for effective and precise communication between scientists across disciplines. Here we provide a list of terms that we will use throughout this paper and their definitions. By condensing the vocabulary and providing precise definitions of often abstract concepts, we hope to operationalize the study of the structure and behavior of complex systems.

- **Unit, element, or node:** an individual object, agent, or part of a system. Unless otherwise specified, we denote the set of n nodes by $V = \{v_1, v_2, \ldots, v_n\}$.
- **Relation:** a set r of one or more nodes, such that $r \subseteq V$. In practice, node relations can arise from correlations in data, observed interactions between units, or groups of elements known to function collectively. A relation r can be *dyadic* if it contains exactly two units ($|r| = 2$), or *polyadic* if the relation contains three or more units ($|r| > 2$). If r contains k nodes, then we say *the k nodes in r are related*. In some parts of the literature, polyadic relations have also been called "higher order" relations, and have been used to refer to motifs in graphs [24]. To avoid confusion, however, in this paper we will use "higher order" to refer exclusively to a particular formalism introduced in Section 3.5. We denote the set of relations by $\mathscr{R}$ unless a domain-specific convention already exists.
- **Property:** information attached to a node or relation. We call the set of properties $\mathscr{P}$ and let p be the assignment map sending $V \times \mathscr{R} \to \mathscr{P}$. For example, a relation formed by the co-firing of neurons can be assigned a frequency, and a relation formed among individuals can have a categorical property such as "teammates". In this work we focus on the units and relations in a complex system, as these are common to all complex systems. Additional properties, including dynamics, are also crucial for system function, but our scope is limited to the structural representation of complex systems.
- **System:** a collection of units V, relations $\mathscr{R}$, and (optionally) any properties $\mathscr{P}$, such that the collection needs no other pieces in order to function completely or to interact autonomously with its environment. The set of units are the components of the system, while the patterns found in the set of relations are called the system's *structure*. An example of a such pattern would be finding a particular node involved in far more relations than expected. The system's activity, including changes in nodes, relations and properties over time, is sometimes called its *function* or *behavior*. An example of behavior would be finding that a the number of relations a particular node is involved in fluctuates over time.

- **Complex system:** a system whose units and relations together exhibit a qualitatively different functionality than the sum of its units acting individually; the main object of study. In this work, "system" always refers to a complex system.

- **System fragment:** a subset of the nodes and relations of a system. Formally, if we write a system as a tuple of nodes and relations $(V, \mathscr{R})$, a system fragment would be written $(V', \mathscr{R}')$ with $V' \subseteq V$ and $\mathscr{R}' \subseteq \mathscr{R}$ a set of relations on node set V'. Researchers usually do not have access to all units or all relevant relations. Instead, they usually have access to – and must perform their studies on – fragments of a system. Sometimes this limited access is due to the vast number of units (a human brain contains on the order of 10^{11} neurons); other times it is due to the inability of our current tools to record all the relations among them (genes that express at low levels are difficult to detect); still other times it is due to other constraints (social media companies may not release their data due to privacy concerns). We do not require a system fragment to itself operate as a system; that is, a system fragment may not necessarily have the ability to fully function or interact with its environment. Consider the complex system of cell metabolism in humans. Even with contemporary tools, we do not have access to all data pertaining to this system. In order to study it, we usually focus on a single aspect most relevant to the question at hand; for example, the set of all experimentally quantifiable proteins (units) and the set of known protein complexes that they form (relations). We refer to the combination of these two sets as the "protein complex fragment" of the cell metabolism system.

- **Dependency:** a property of a system in which the existence of one relation provides information about the existence of another relation. In this case we could say one relation is dependent on another relation. Conversely a relation is independent from another relation if the existence of one relation in no way affects the (probability of the) existence of the other. See Section 2 for formal definitions of the three types of dependencies we discuss in this work.

- **Formalism:** a mathematical framework or theory (a collection of definitions, results, and theorems) that can be used to represent, model, encode and study a complex system. In this paper, we will explicitly discuss the graph, simplicial complex, and hypergraph formalisms.

- **Representation:** a mathematical or computational encoding of a specific complex system (or a fragment of one). A representation is the materialization of a specific formalism, e.g. it is one concrete, specific graph, as opposed to the mathematical theory, or formalism, of graphs.[1] For example, one might study the brain by *representing* it as a graph with a node for each lobe and edges joining two nodes if they are physically adjacent. In this case, the brain is the system, graph theory is the formalism, and the graph of n nodes that mirrors the brain connections is the representation.

- **Encode:** the process of taking a system or data collected from a system and formulating it as a representation using a specific formalism.

In the rest of the paper we will assume the reader has already defined what should constitute a node and relation within their system. We refer the reader to [37] for a thorough discussion regarding how to choose nodes and relations when these choices are not straightforward.

2 Dependencies by the system, for the system

When studying or modeling a complex system composed of many parts, several design decisions must be made. We begin by considering one specific and rather fundamental choice, which is sometimes only implied and other times outright neglected. This choice regards the decision of which *system dependencies* one should seek to appropriately and accurately encode. Reiterating our definition above, *a dependency is a property of the system in which the existence of one relation provides information about the existence of another*

[1] For readers familiar with object-oriented programming, we liken the difference between "formalism" and "representation" to that between "class" and "object".

relation. Said another way, does the system have underlying rules or restrictions that cause interactions to occur or units to behave in particular ways? For example in a social system of individuals and friendships, if two individuals live physically close to one another, then their likelihood of becoming friends is larger than if they lived far apart. Furthermore, if they live near each other, then they are also more likely to meet and consequently befriend each other's neighbors. In this way, knowledge of the existence of one friendship informs us of the possible existence of other friendships, because the friendships (relations) between people (units) are affected by geographical distance (dependency).

Such system-level dependencies can manifest in different ways; here we will constrain ourselves to a discussion of three of the most commonly observed dependency types. Specifically we discuss *subset* dependencies (does a large relation influence the existence of smaller sub-relations?), *temporal* dependencies (does temporal nearness of elements influence their relations?), and *spatial* dependencies (does the physical proximity of elements influence their relations?). We acknowledge that dependencies other than those described in this work exist within real-world systems; in many domains of inquiry, ongoing research efforts seek to define the proper avenues for illuminating dependencies and approaches for their incorporation.

2.1 Subset dependencies

When investigating a complex system, we often record its elements and the observed relations containing two or more of those elements. For example, we might record objects and shared observable features [128], people and shared conversations [229], or neurons and their co-firing [53]. Here, we can think of the system as a set of nodes V and a set of observed relations $\mathscr{R}$ in which each relation $r \in \mathscr{R}$ is a subset of V and is meant to represent one observed interaction between k elements. In this setup, some nodes may participate in many relations, while others participate in very few or none at all. It is then important to ask: if we observe the relation $r = \{v_0, \ldots, v_{k-1}\} \in \mathscr{R}$, does it imply that some subset r' of r is also a relation? If so, the system exhibits the type of dependency that we call a *subset dependency*. For example, in the words-and-features system fragment, if three words (ball, egg, globe, written as v_1, v_2, v_3) correspond to objects that share a particular feature (each of them is round, so that 'is round' defines a relation $r = \{v_0, v_1, v_2\}$), then any two of the objects must also share that same feature (then $r' = \{v_0, v_1\}, r'' = \{v_1, v_2\}$, and $r''' = \{v_0, v_2\}$ are all relations). One can make a similar argument for people conversing with one another and for neurons co-firing. In these cases, every subset of any set of related nodes is also related. However, we will see examples later when only some, or none, of the relation subsets are also relations, and we will describe this scenario as indicating the presence of a different type of dependency. Concretely, *we will say that a system with nodes V and relations $\mathscr{R}$ exhibits a subset dependency if for $r \in \mathscr{R}$ and $r' \subset r$, we must have that $r' \in \mathscr{R}$ whenever $P(r')$ is true, where P is some logical predicate.* For instance, in the words-and-features system, the logical predicate determines whether words corresponding to objects share a feature. In that system, since a subset of words for objects in a relation always share a feature, the logical predicate is always true, and we see clearly that a subset dependency exists in the system.

To illustrate this specific type of dependency, in Fig. 2 we show a system fragment of chemical reactions (left) and a system fragment of objects with shared physical descriptors (right). On the left side of Fig. 2, molecules or compounds correspond to nodes, and reactions define relations between nodes so that if k compounds together exclusively form the reactants and products of one reaction, then those k nodes are related. We see that O_2 and H_2O participate in multiple reactions together, for example $2H_2 + O_2 \rightarrow 2H_2O$, but we do not observe a reaction that *exclusively* uses O_2 and H_2O. Therefore this system fragment does not display the property that all subsets of relations are also relations, since we have that $\{O_2, H_2O\} \subset \{H_2, O_2, H_2O\}$ and $\{H_2, O_2, H_2O\} \in \mathscr{R}$, but that $\{O_2, H_2O\} \notin \mathscr{R}$. In contrast, the right side of Fig. 2 shows a collection of objects and features (shape and color), in which each object may share physical features with other objects. In this case a relation $r_{\square} = \{■, ■, ■\}$ contains all objects that are square. Notice that by our definition of relation for this system fragment, we immediately get that $r' = \{■, ■\}$ is also a relation. Specifically, the pink and red squares are related because they share the feature "square", but also any subset of the squares will also be related because they, too, share the feature "square". This example of objects and shared features does display the subset dependency, since subsets of related nodes are also related.

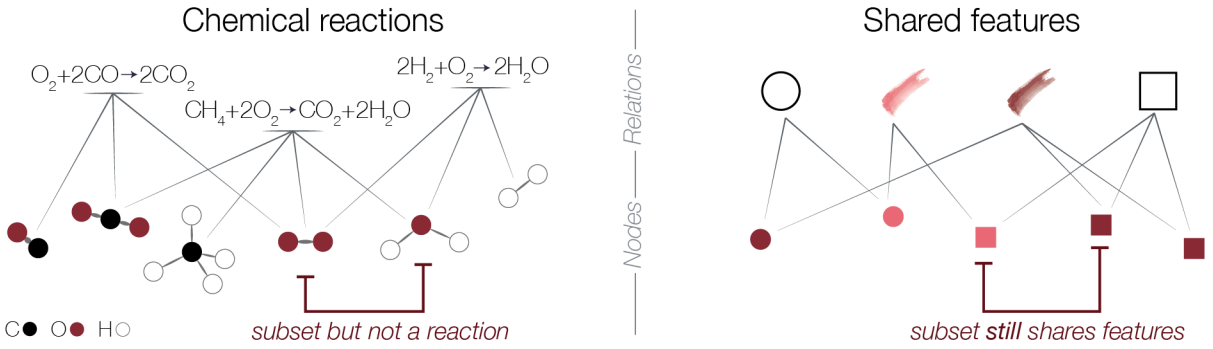

Figure 2: **Are subsets of related nodes necessarily related?** Systems may exhibit a *subset dependence*, which occurs when a relation between nodes implies the existence of a relation between any subset that satisfies a certain logical predicate. *(Left)* System fragment composed of molecules and chemical reactions. Here, O_2, H_2O, and H_2 participate in the reaction $O_2 + 2H_2 \rightarrow 2H_2O$, but a subset of these compounds does not independently engage in a reaction, such as O_2 and H_2O. *(Right)* System fragment composed of objects with observable features such as color and shape. All objects that are squares are related by the presence of the shared feature "square". Any subset of these square objects will also still possess the shared feature "square", and thus will also be related. In this case, the logical predicate is always true.

When a system displays a subset dependency, we must ask ourselves whether we should explicitly represent that property in our model. The answer to that question will depend on, among other things, the available data, the research question, and how we define relations among nodes. Incorporating the subset dependency in a representation usually requires the data to include polyadic relations, which are not always directly observable. Additionally if the research question involves trajectories through related nodes, it may or may not be necessary to incorporate polyadic relations and thus the system's subset dependencies explicitly, since often we can answer questions about trajectories between nodes using exclusively dyadic relations between nodes.

Most commonly, the choice of whether to include the subset dependency affects the formal representation used to encode the system, and consequently the results of downstream analyses. For example, if Marta is involved in a group of people having conversations and we define relations as shared conversations (so that a subset dependency exists), then if we count the number p of people with whom Marta converses we do not know if Marta had p separate conversations with each of the p individuals, or if she participated in one large conversation with all p people. Without a distinction, Marta's popularity with others could be vastly over- or under-estimated. This example illustrates how the occurrence of subset dependence can be determined by the definition of relation. In Section 3 we explore the benefits and drawbacks of a few abstract formalisms that capture different types of dependencies. For now, we stress that the presence or absence of subset dependencies influences the computations we can perform and the formalisms we can use.

2.2 Temporal dependencies

Next we consider systems in which we observe information, individuals, or goods moving along trajectories through time. A simple example would be a city subway system where passengers ride the train from one stop to the next until they reach their destination. In such systems we must ask the question: Does the current location of an individual affect where they might move next? *We say a system exhibits a temporal dependency if the existence of relations at time t affects the behavior of units or relations at time $t' > t$.* Said another way, it may be that trajectories or walks within systems that display temporal dependency are not Markovian, since the future trajectory of a walker depends not only on its current location but also on some previous trajectories of itself or other units.

Consider a subway system in which passengers can travel via trains to stations A through H (Fig. 3). If

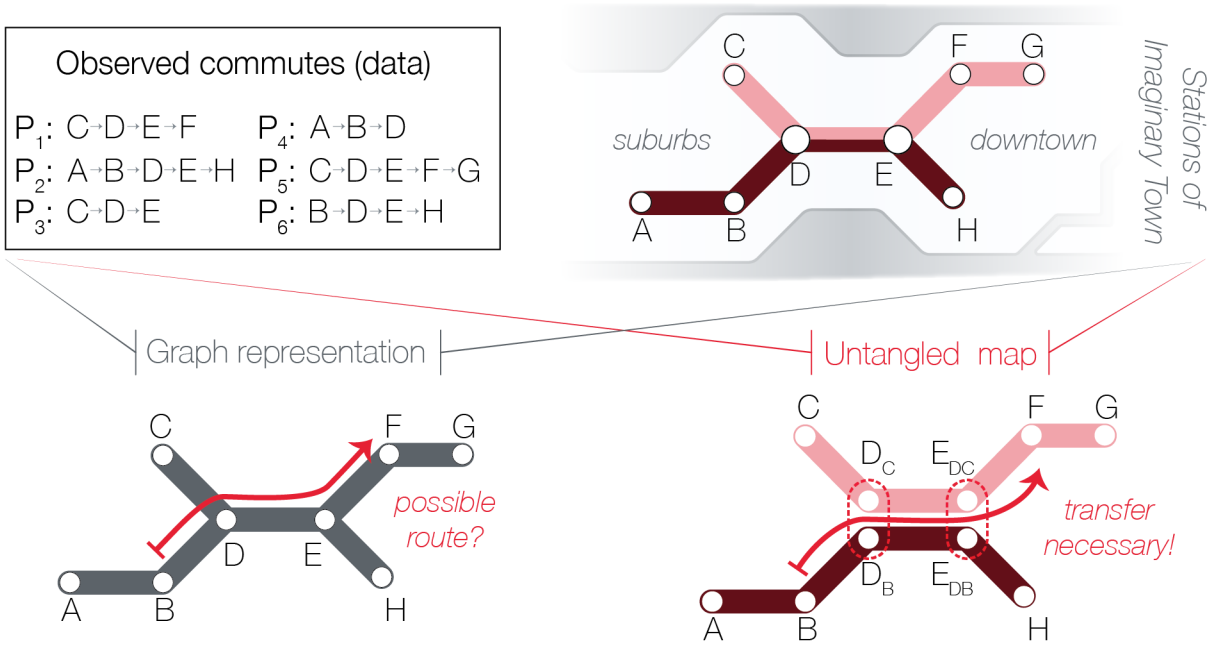

Figure 3: **By incorporating temporal dependencies into the representation, we obtain a more accurate subway map.** Given data from six commuting passengers (P_1, P_2, ..., P_6) who do not switch trains *(top left)*, how can we obtain the underlying subway map? We could create a graph in which two stations are connected if a passenger transferred from one station to another. However, such a graph would suggest that a passenger could commute from station B to F without switching trains *(bottom left)*, which is not possible in this system. If instead we untangle the subway lines by respecting the temporal dependency and treating trains that arrive to station D from station C as different from those arriving from station B, then we can clearly see the necessary transfer between subway lines required for the B to F commute *(bottom right)*.

our complex system consists of passengers commuting via the subway, then our observed data might include explicit passenger routes. For example, in Fig. 3 we record the routes of six passengers, each of whom commutes from the suburbs (stations A, B, and C) to downtown (stations D, E, F, G, and H). For the purpose of the example, we assume that passengers do not transfer between distinct train lines during their commute. If we now represent our data as a set of stations (units) and we connect two stations i and j if j immediately follows station i in at least one passenger route (relations), we obtain the subway map shown in the bottom left of Fig. 3. Because this diagram records all known movements of passengers between pairs of stations, we might confidently proceed to the next analysis step. However, it is worth noting that this particular representation suggests that the red path from station B to station F is a possible commute for a passenger. Yet when we look back at the data itself, such a commute seems extremely unlikely since the sequence $B - D - E - F$ never occurs. The fact that this route appeared natural from the representation, but not from the data, points to the fact that our system contains a temporal dependency and, importantly, that this dependency is not well reflected in the particular representation we chose.

As discussed in great detail in [24, 174, 69, 59, 118, 158], the fundamental limitation of keeping only pairwise sequential relations, as done in the bottom left of Figure 3, is that in the representation we assume that traversal across each link is Markovian and therefore its probability is independent of the probability of traversing any other link in the system. More explicitly, paraphrased from [117], by representing the system as a graph (see Section 3 for a definition) we assume that the edges (i, j) and (j, k) are independent and

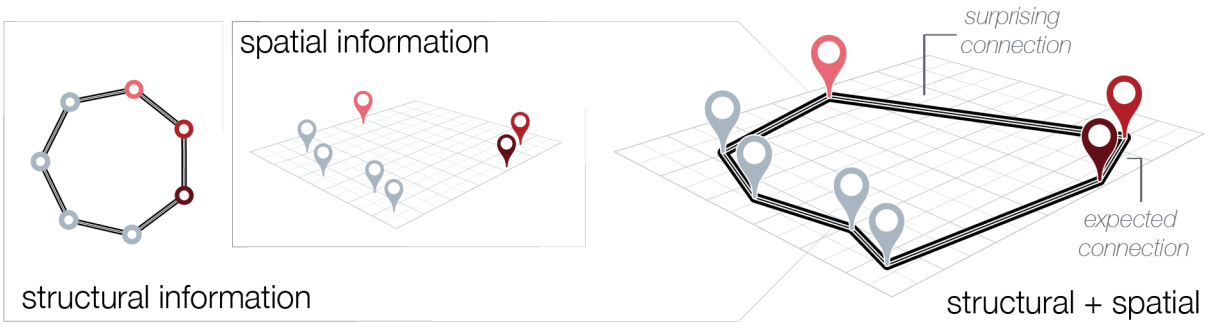

Figure 4: **Spatial dependencies within a system can complicate our representations of the data.** In our example system, we have *(Left)* connection information that is independent of any system embedding, and *(Middle)* spatial information indicating where the nodes physically reside. A possible combination of the two information types *(Right)* can be used to better understand the physical constraints on the topology. If long distance connections are costly for that system, the combined representation allows the investigator to assess the prevalence and location of those costly (and thus potentially surprising) connections.

that the two-step transition from i to k proceeds in two independent steps. This assumption can easily be violated by a real system, as seen in our toy example, since sometimes one step in this traversal is dependent on which steps came before (i.e. transitions are not Markovian). Mismanaging temporal dependencies in systems can lead to misleading results that can, for example, over-represent the importance of edges rarely used or create non-existent connections. We will discuss a formalism that is particularly appropriate for representing temporal dependencies in Section 3.

2.3 Spatial dependencies

The third and final type of dependency that we discuss here arises from the physical nearness of units within a system. For example, in the human connectome a brain region is likely to extend white matter tracts to neighboring regions, providing physical conduits for electrical activity [200]. In granular materials, resistance to external forces relies on interactions between only particles that physically touch [156]. More generally, many spatial systems are so named because the spatial location of nodes affects their likelihood of interacting with one another [13, 14]. *Here we say that a system exhibits a spatial dependency if the distance between two or more nodes influences the existence of a relation that contains them.* More formally, consider a system whose nodes are labeled by $V = v_1, v_2, \ldots, v_n$ and each node v_i has associated to it a point x_i in some metric space. Then, this system exhibits a spatial dependency if the probability of a relation between nodes $v_1, v_2, \ldots, v_k$ is a function of the pairwise distances between the corresponding locations $x_1, x_2, \ldots, x_k$.

Many such systems exist in the natural and manufactured world. Indeed, spatial restrictions influence communication in cell populations [116, 165], trade in economic networks [98], and passengers in transportation networks [221, 122]. As an example of spatial dependency within an abstract system, we might begin with only knowledge of the pattern of related nodes. We display this structural information in the left panel of Fig. 4 with circles corresponding to nodes and lines joining circle pairs whose corresponding nodes are related. From the structural information alone we might expect that relating the pink and red nodes is just as difficult or costly as relating the red and dark red nodes; we might therefore infer that the two relations are equally crucial to the system's function. However, if the system exists within an environment containing coordinates and a distance function, with each node having spatial coordinates and a measure of distance between each pair, then this spatial information could offer a different perspective on the system. In the middle panel of Fig. 4, we see that the nodes, now depicted with colored pins, are spread out so that some are more spatially clustered whereas others are less so. Considered alone, the spatial information gives us no insight into the actual relations present in the system, but does provide information with which we might

predict the likelihood that nodes are related.

In many spatial systems such as the brain or city transportation, relations between distant nodes are unfavorable due to a higher cost of creation and maintenance, while short-range relations are far easier to construct. In the face of this association between the physical distance across a relation and its cost, we might consider the distances between nodes and infer that the red and dark red nodes are likely to be related, while the pink and red nodes are not. When we finally combine the topological and spatial information (Fig. 4, right), we then can leverage the two information types to understand which relations are most surprising or make hypotheses about which relations are most important to the system. For example, the dyadic relation between the pink and red nodes might be very costly given the long distance, so we might infer that the pink to red relation is more essential to the system than the red to dark red relation since the system would only spend valuable resources to maintain such a relation if it was integral to system function. Without the spatial information, we may have incorrectly placed the same importance on the pink-to-red and the red-to-dark red relations. This example highlights one of many ways in which we could integrate spatial and structural information.

As with the previous dependency types, failure to account for a spatial dependency can greatly bias our models and results. Consider an outbreak of a contagious disease. If we recorded the habits of infected individuals such as their diet, but fail to record their locations and physical mobility through space [208, 7], then we might – for example – wrongly attribute disease spread to the broad consumption of a particular food that is prevalent in the infected region instead of through person-to-person contact. As another example, social contacts are also influenced by proximity. If we return to evaluating Marta's popularity, the observation that she has many friends may come from the fact that she lives in a densely populated area, rather than from her charisma or personality. In these examples, failing to account for spatial dependencies may result in attributing certain structural properties of the system to the wrong cause.

2.4 External sources of dependencies

Before we shift our focus to concrete ways of encoding system dependencies using mathematical formalisms (Section 3), it is useful and interesting to consider how external forces can influence the observed system dependencies. Ideally, we as investigators would have the ability to measure all dependencies within the system under study, and then use this knowledge to make an informed decision as to the appropriate formalism with which to model our system. However, often the processes of scientific inquiry do not proceed so effortlessly: no analysis is ever devoid of the influence of external factors, or biases. Our goal in this section is to highlight possible sources of such bias. Although we have already discussed biases arising from dependencies native to the system under study, here we emphasize that acknowledging and understanding dependencies imposed by outside sources should also play a crucial role in determining an appropriate representation and subsequent analyses.

- **Data availability.** One notable and common constraint in science is the limited data that can be empirically acquired from a given system. In other words, researchers usually have access only to a fragment of the system. As a consequence, any dependency that is observed and ultimately encoded may be determined more by the sparsity of available data than by the system's true structure and function. For example, one may have access to only sparse snapshots of or short sequences from an evolving system [188], making the subset dependency difficult to identify and effectively encode. Particularly, there may not be enough data available to correctly deduce the predicates P that a subset must satisfy in order to also form a relation (see the definition of subset dependency in Section 2.1).

- **Data acquisition or processing.** Certain experimental techniques or computational procedures may produce spurious dependencies. A common example involves correlation matrices. By computing the correlations of node activity (a common approach in fMRI-based functional connectivity matrices [211, 88]) one induces a transitivity dependency, which is a type of subset dependency. Concretely, if A, B, C are nodes in a system where two nodes are related if the time series of their activities are highly correlated to each other, as determined by some data acquisition method, then whenever A and B are

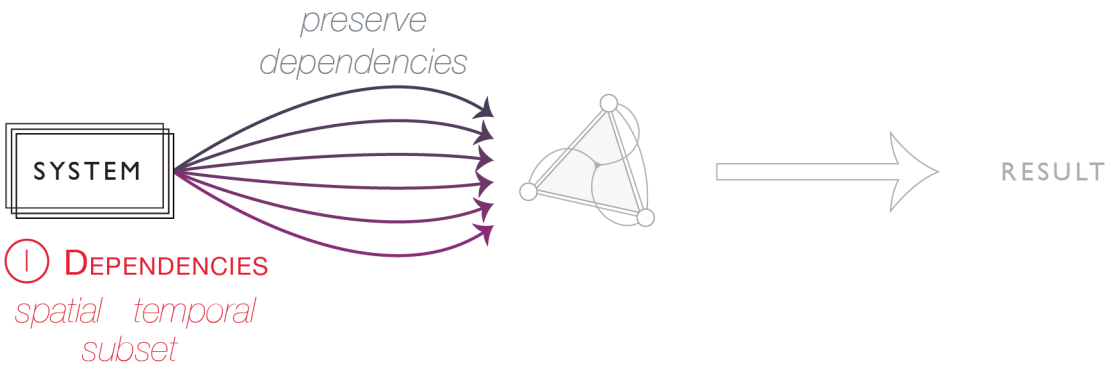

Figure 5: **Understanding system dependencies is a first step in the complex system analysis pipeline.** Types of dependencies include spatial, temporal, and subset dependencies. Acknowledging dependencies at this step allows for proper preservation of dependencies throughout the rest of the analysis pipeline. When preserving all dependencies is not possible due to factors outside the control of the researcher, acknowledging this inability frames the results in a proper context.

related, and B and C are related, it is highly likely that A and C are also related. In this case, it is possible that relations between nodes implied by the calculated correlations are found in the processed data but not in the system itself. For example, one might find that changing the type of correlation results in a change in the inferred relations.

- **Research question.** The research question at hand will influence which relations within a system are particularly interesting. Moreover, it may also influence the very definition of a relation. For example, consider a system of proteins that interact to form protein complexes. If we wish to study which proteins appear together in many complexes, then we may define a relation as k proteins that participate in the same complex. If instead we wish to study protein complexes themselves, we could define a relation as a set of k proteins that all together form a single complex. In the first case, the relations are tied to a subset dependency (if three proteins appear together in a complex, then so do any two of them), but the second does not. On the flip side, a given research question may neglect a relevant dependency in the system. For example, we could ask if a common food could have caused a disease outbreak. Answering that explicit question neglects the fact that individuals near each other will likely eat similar foods. The research question is not broad enough to incorporate the spatial information as part of the answer, and therefore spatial dependencies may seem irrelevant at first sight, when they may be in fact essential to finding the real answer. We expand upon this topic in Section 3.5.

To summarize, we have defined and discussed three types of dependencies that could exist in a complex system: subset, temporal, and spatial. We emphasize that dependencies can arise from within the system itself or from external factors, but regardless of their origin, we as researchers must be aware of their existence and how they influence our models and results, especially given their early position in our analysis pipeline (Fig. 5). As we will continue to see in the sections that follow, the recognition and encoding of dependencies can greatly affect the results of our analyses and the conclusions that can be drawn.

3 Formal representations of complex systems

Over the years many representations of complex systems coming from different mathematical and computational formalisms have taken hold across scientific disciplines. Different formalisms allow for the modeling of unique aspects and dependencies of each system, but the multiplicity of available formalisms presents challenges for the communication, collaboration, and ultimately the progress of complexity science. Furthermore, the choice of formalism also complicates the analysis pipeline that researchers must decide upon when studying a particular system.

Here we discuss three of the many possible mathematical formalisms that researchers commonly use to represent their system: graphs, simplicial complexes, and hypergraphs, chosen for their prevalence in the complex systems literature. A complex system is, at its core, a collection of units and their relations, therefore we require our representations to mirror this composition of units and relations. The units of all three formalisms discussed here are called *nodes*. *Graphs* represent pairwise relations among nodes as *edges*. Despite their simplicity (or perhaps because of it), graph representations have supported several important discoveries such as the prevalence of small-worldness in real-world networks [220, 6]. Still, graphs can only, by nature, represent dyadic relations between nodes[2]. If instead relations within the system exist between more than two nodes, one might turn to either a *simplicial complex* or a *hypergraph*. Both of these formalisms naturally allow us to encode such polyadic relations [21]. The relations represented by a simplicial complex are called *simplices* and those represented by a hypergraph are called *hyperedges*. We will first define each formalism, so that later in this exposition we can explicitly discuss their respective advantages and assumptions.

3.1 Graphs

The first and perhaps most common formalism used to model complex systems stems from graph theory. A *graph* G is a collection of nodes and edges between nodes such that an edge connects at most two nodes (Fig. 6, left). We denote the set of nodes as V and the set of edges $E \subseteq V \times V$, so that a graph is defined uniquely by $G = (V, E)$; note that each edge is an unordered set of two nodes. The nodes of a graph are the units, and edges describe how these units are related. If v_A and v_B are nodes of the graph, then we write (v_A, v_B), or $v_A - v_B$ to represent the fact that the two nodes are connected by an edge. Studies that form a graph representation from the underlying data frequently involve finding densely connected sets of nodes or determining how an object might traverse the structure. In using the graph representation, such questions could lead to detecting *cliques* or *communities* in the graph, or identifying chains of connected nodes called *paths* (see Fig. 6, left, and Section 5.1 for more examples).

Many attribute the origin of graph theory to Leonhard Euler in the 18th century [74]. One can also trace its presence outside of mathematics back to the use of sociograms and social network analysis in the 1930s [78], and to graph-like data structures in computer science in the 1950s [222]. Notably, the use of graphs to model more general complex systems has rapidly increased over the past few decades, driven largely by the discovery of the small-world effect [220] and heavy-tail degree distributions [10] in real-world datasets. Encoding a system as a graph has the great advantage of hundreds of years of mathematical theory behind concepts, generally simple computations, and insightful visualization. However, the graph by definition assumes that relations between nodes occur exclusively at the pairwise level. Systems such as transportation networks might solely contain pairwise relations among their units, but many others, especially from biology, often have polyadic relations. Still, the graph's ability to model systems has proven quite useful in distinct fields such as neuroscience [17, 36], computer science [75, 132], and ecology [161, 135].

[2]More precisely, edges in a graph can only involve (at most) two different nodes. Whether the interpretation of each of those nodes is that of a single unit or many units (as is done for example in some representations that involve the idea of a "supernode"), is not a relevant matter for graph theory, but for the process of encoding data into a graph.

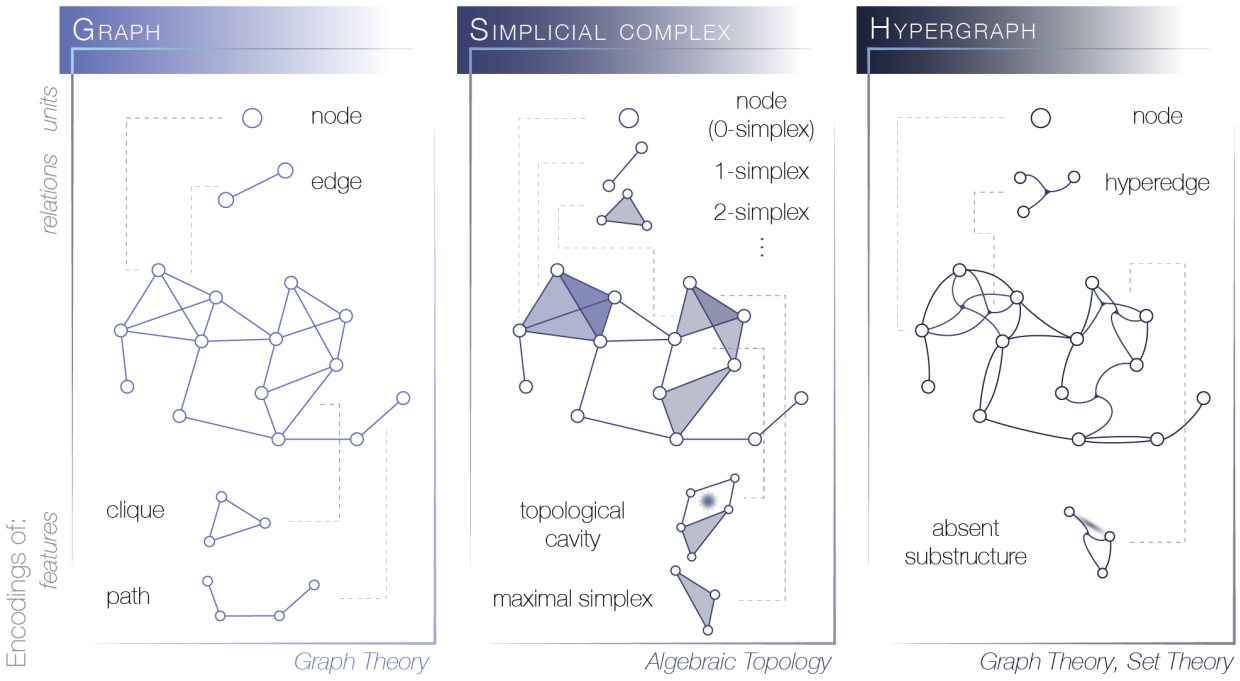

Figure 6: **Three types of formalisms composed from nodes and relations**. *(Left)* Graphs involve units called nodes and relations between two nodes called edges. Possible features of interest for graphs include all-to-all connected sets of nodes called cliques, as well as routes between nodes called paths. *(Middle)* Simplicial complexes can be used to represent systems with polyadic relations among units. Sets of related nodes are connected by simplices. A k-simplex describes $k+1$ nodes that collectively interact, such that any subset of nodes forming a simplex must also form a simplex; this is called "downward inclusion". Motifs of interest include topological cavities and maximal simplices. *(Right)* Hypergraphs can also be used to represent systems with polyadic relations among units. Sets of related nodes are connected by hyperedges. Hypergraphs are not restricted by downward inclusion. Of particular interest within a hypergraph is the absence of a substructure (or smaller relation), for example in which two nodes do not connect dyadically but participate together in a hyperedge that connects a superset of the node pair.

3.2 Simplicial Complexes

The next formalism that we consider addresses the need to acknowledge polyadic relations in the system. Illustrated in the middle column of Fig. 6, a *simplicial complex* is a set of nodes V (also called *vertices* in the field) along with a collection of subsets of nodes $\mathscr{R}$ (often denoted by K in the field) such that for any $r \in \mathscr{R}$ and $r' \subset r$, we have $r' \in \mathscr{R}$; we will refer to this condition as "downward closure". A set of $k+1$ nodes $r \in \mathscr{R}$ is called a k-simplex, and downward closure requires that any subset of nodes within a simplex also forms a simplex. In practice we often imagine a k-simplex to indicate an application-relevant interaction between the $k+1$ nodes, such that these nodes may function in unison. The simplicial complex (precisely, the *abstract simplicial complex*) would then record the individual units (nodes), the functional building blocks (simplices), and how all these building blocks are assembled into one system (the simplicial complex). Since subsets of simplices are simplices by definition, then if k nodes are related, we have that any subset of those k nodes are also related. The simplicial complex can be written as a binary *incidence* matrix of dimensions $\#maximal\ simplices \times \#vertices$ where an element containing a 1 indicates node participation in the corresponding maximal simplex; a maximal simplex is a simplex that is not contained in any larger simplex.

Although algebraic topology has been studied for well over a century, it was not until the late 1990's that applied algebraic topology as a discipline began to emerge [230, 68] (though we note a earlier uses exist [9]). Many of the earliest studies used applied topology and simplicial complexes to study data in the form of point clouds [40, 189]. Later, it became clear that the simplicial complex language was a natural formalism for explicitly representing biological and physical systems. For example, simplicial complexes have been used to represent neural recordings [86, 53], classify images [203, 55, 66], and describe the mesoscale architecture of brain networks [201, 202, 167, 191, 159]. Even more recent work has focused on defining generative models to construct simplicial complexes with given topological features [51].

3.3 Hypergraphs

The final formalism that we consider draws again from sets of nodes and their relations, yet is even more general than the simplicial complex discussed above. The *hypergraph* is an extension of the mathematical definition of a graph, in which we have a node set V and a hyperedge set $\mathscr{R}$ (sometimes denoted in the field as $\mathscr{E}$). A hyperedge $e \in \mathscr{R}$ can connect an arbitrary number of nodes. That is, while an edge in a graph can only connect two nodes, a hyperedge can connect three, four, five, or more nodes (Fig. 6, right). More rigorously, a hypergraph is a pair $(V, \mathscr{R})$ with V a node set and $\mathscr{R}$ a set of subsets of V [214, 26]. In contrast to the simplicial complex, we can use the hypergraph to encode polyadic relations without the restriction of downward inclusion. Formally, a subset e' of a hyperedge e, $e' \subset e \in \mathscr{R}$, does not necessarily exist as a hyperedge. Additionally, we can rewrite a hypergraph as a binary *incidence* matrix of dimensions $\#hyperedges \times \#vertices$ in which an entry of 1 indicates the node participation in the hyperedge.

As noted above, the crucial restriction that is relaxed when moving from a simplicial complex to a hypergraph is that of downward closure. Recall that in a simplicial complex any subset $r' \subseteq r$ of a simplex r must also be a simplex. Hypergraphs do not obey this rule. For example we may see a hyperedge connecting vertices v_1, v_2, and v_3 but no hyperedge that connects v_1 to v_2 exclusively. Or, given two hyperedges connecting nodes v_1, v_2, v_3, and v_2, v_3, v_4, if a hyperedge connecting v_2, v_3 also existed, does this smaller hyperedge indicate a sub-relation for the hyperedge v_1, v_2, v_3, the hyperedge between v_2, v_3, v_4, neither, or both? With a hypergraph, we cannot determine how or if a sub-relation emerges due to superset relations (see [195] for a deeper discussion). This subtle difference allows hypergraphs to represent a wide diversity of systems, including many that the simplicial complex formalism would not appropriately represent. The hypergraph's increase in modeling flexibility is counterbalanced by a decrease in formal analysis methods, which we will discuss more in Section 5.

The flexibility and ability to model polyadic relations made hypergraphs an appealing formalism in many systems that were originally studied with graph theory. Indeed one of the earliest practical uses of hypergraphs was to understand social networks [184]. Since then, researchers have successfully employed hypergraphs to study polyadic relations in the Enron email dataset [162], find the core of yeast protein-protein

interactions [163], uncover motifs in neurodevelopment [89], track changes in evolving systems [18, 56, 57], and detect failure in biochemical networks [110]. As many uses of hypergraphs arose out of systems first modeled with graphs, many analysis methods for hypergraphs mimic those originally used for graphs (we discuss this point further in Section 5.3).

3.4 Variations

We note that the above descriptions only scratch the surface of complex system encoding possibilities. An ever broadening set of scientific questions drives the need for novel variations of each formalism, resulting in a myriad of definitions and manipulable parameters. One could extend our mathematical definition of complex systems to include the following properties, as a map $p : V \times R \to \mathscr{P}$ where $\mathscr{P}$ is a set of properties we care about, as mentioned in Section 1.1. Here we note a few of the most common modifications to each of the above formalisms, driven by the need to incorporate more information about the system at hand.

Directed

Many complex systems including the brain, transportation networks, and metabolic pathways exhibit directionality in their relations. That is, in these systems, if v_A and v_B are units that share a dyadic relation, there is a meaningful distinction between a relation where v_A comes first, one where v_B comes first, and one where either v_A or v_B comes first (but there must always be an order in how they are related). To distinguish these cases we write $v_A \to v_B$, $v_B \to v_A$, or $v_A \leftrightarrow v_B$, respectively. If we apply this idea to the graph formalism, a *directed graph* is one where each edge is now an ordered set of two nodes. Directed graphs have proven extremely useful in many contexts from scheduling and monitoring workflows [112, 1] to cardiac excitation modeling [212] to understanding percolation processes relevant to wild fires and other explosive phenomena [197, 65]. Moving to simplicial complexes, directionality is still quite natural. Indeed simplices themselves inherit a directionality, formally known as an *orientation*, encoded by the numbering of the participating vertices. In practice, in an oriented k-simplex, each node is made to point only to nodes with a higher assigned number. Oriented simplicial complexes arise in practice from directed synapses between neurons [167] as well as directed migration flow [100]. Finally, in hypergraphs, one may represent directionality with *hyperarcs*, the term for a directed hyperedge. More formally, a hyperarc is a pair of disjoint subsets of vertices with one subset comprising the sources and the other subset comprising the sinks [82]. Directed hypergraphs have proven useful in constructing a biological pathway database [113], tackling problems in computer science such as propositional logic [82] and combinatorial optimization [119, 87], and finding specific patterns of connectivity in chemical reaction systems [149], among others.

Weighted

In real-world systems, not all relations are created equal; even within the same system, relations between individual units may vary in strength or magnitude. To represent these differences, the strength of a relation can be encoded using the *weighted* versions of the above formalisms. To assign weights to any of the above encodings, we can define a general weight function $W : \mathscr{R} \to \mathbb{R}$ from the set of relations $\mathscr{R}$ (edges, simplices, or hyperedges) to the real numbers $\mathbb{R}$. For a graph, this function would assign a value to each edge, which we generally interpret as the strength or frequency of the pairwise interactions between the corresponding nodes. In the context of weighted representations, the original versions containing no weights are called *binary* or *unweighted*, as they can be cast as weighted objects where the weights of all relations are either one, if they exist, or zero if they do not exist. The brain connectome, traffic between municipalities [62], and functional similarity of genes [153] have all been modeled as weighted graphs. Additionally, many common graph metrics such as the clustering coefficient and path length (covered in more detail in the next section), extend easily to the case of weighted graphs [175], making this variant of representation particularly pervasive. Similarly we can construct a weighted simplicial complex by assigning a weight to each simplex. However, recall that in a simplicial complex any face of a simplex must also be a simplex, and thus if we have a relation between k nodes then any subset of these nodes must be related to at least the same extent

as the superset. Said another way, we require that the weighting function W on simplices adheres to the rule that for any simplex r, if $r' \subseteq r$ then $W(r) \leq W(r')$. Weighted simplicial complexes can arise from point clouds with inverse distances between points as weights or from growing processes with time of addition used to assign simplex weight. Perhaps most often, we study weighted simplicial complexes through the lens of persistent homology, which computes the organization of topological cavities housed within the weighted simplicial complex [230, 39, 83, 148] (see a few recent uses in [191, 86, 159, 201]). Lastly, in hypergraphs we can naturally weight hyperedges with distinct values [82]. Importantly, weighting hyperedges allows more flexibility in choosing weights, as weighted hypergraphs do not enforce rules restricting weights on subedges in contrast to weighted simplicial complexes. Weighted hypergraphs have proven useful in image segmentation [169] and in the process of incorporating prior knowledge into learning algorithms [207].

Dynamic

Complex systems such as cell signaling, traffic patterns, and transactional relations also grow, separate, or fluctuate in time [124, 176, 44, 121]. Consequently, formalisms have been adapted to represent such an evolving architecture. A *dynamic graph* or a *temporal graph* is a sequence of graphs $G_1, \ldots, G_T$ in which each G_i is a graph on the same set of nodes, and each node is mapped to its identity when moving from G_i to G_{i+1} [95]. As with other variations on graphs, multiple computational tools such as community detection have been extended to include these types of dynamics [143, 190, 138]. Moving to simplicial complexes, a dynamic simplicial complex is similarly a sequence of simplicial complexes on the same node set. Questions about the topological cavities of simplicial complexes can be answered by using *vineyards* [224] and *zig-zag* persistent homology [131] depending on the types of evolving simplicial complexes. Finally, a dynamic hypergraph is a sequence of hypergraphs $H_1, \ldots, H_T$ on the same node set where hyperedges may change from H_i to H_{i+1}. At the time of writing, we found few examples of applied dynamic hypergraphs, although we note that their visualizations have been studied [210]. Nevertheless, we suggest that this particular variation of hypergraphs could be useful for example in modeling evolving gene interactions, functional relations between brain regions, and the time-varying structure of social groups.

Multilayer

Often the units or relations of a system have types, categories, or classifications that distinguish them. It is sometimes useful to distinguish between these types of relations in our representations, and one way to do so is to use the so-called *multilayer* variations. Generally, multilayer graphs consist of a set of graphs that may (or may not) involve the same nodes; each graph in the set comprises a *layer*. The graph in a given layer contains relations of exactly one type. Consider a human brain in which two regions might show an increase in blood flow either due to coupled neuronal activity or due to interactions involving nearby blood vessels themselves. To encode these two types of relations in a single representation, we could use a multilayer graph with two layers: one encoding the relations between neurons and another encoding relations between blood vessels. We note that when all layers contain the same set of nodes and the only interlayer edges that exist connect nodes to themselves in other layers, the representation is called a *multiplex* graph [31, 193]. We invite the interested reader to visit [108, 29] for more rigorous definitions, and [129, 35, 226] for implications for diffusion and control. Time-evolving systems can be seen as a subtype of multilayer systems, in which the layers are a set of graphs ordered in time. Previous studies have used multilayer networks to model complex spreading processes [58, 178, 179, 209], understand explosive word learning [198], and uncover the community structure of trade relations [11]. Multilayer simplicial complexes or hypergraphs would similarly include a set of simplicial complexes (respectively, hypergraphs) not necessarily defined on the same nodes in each layer. As of the time of this writing, we did not find applications yet of this extension. We suggest that these variations could prove useful for understanding multiple types of biological data collected on a set of nodes. As an example, one could encode common properties (mutation status, chromatin rearrangements, etc.) as layers in a multiplex network of cancer cell lines in order to better understand drug response [166]. The multilayer variation is readily applicable whenever researchers have access to and want to model two different fragments of the same system.

Higher Order Networks

Higher Order Networks (HONs) are a variation of the graph formalism that aims to represent a certain kind of temporal polyadic relations. Instead of encoding system units as nodes, the HON encodes frequent paths or transitions in the data as nodes, which then allows us to interpret the final representation with the standard Markovian assumptions on edge sequences. Recall our example of commuting passengers in Figure 3. We can build a HON from the observed path data to encode the observed dynamics and temporal dependencies of this system in a particular kind of graph. In Figure 3, the more accurate subway map on the bottom right, reconstructed from the observed data, contains two nodes that correspond to the physical station D. The one labeled D_B represents the passengers that arrive to D from station B, while D_C corresponds to those that arrive from station C. Similarly, the physical station E splits into two nodes: E_{DC} and E_{DB}. The nodes on this map do not correspond to the stations observed in the town's transportation system, but to the possible passenger pathways through them. Indeed, as observed before, we never observe a passenger commute that traces the path $C-D-E-H$: all passengers that pass through stations $C-D-E$, in that order, then go on to station F, while all passengers that pass through stations $B-D-E$, in that order, go on to station H. Therefore, the representation on the bottom right of Fig. 3, an example of a higher-order network or HON, is a more faithful representation of the observed data and its temporal dependency. Note that if the observed passenger data changed to include a route visiting stations $C-D-E-H$, the structure of the HON would change, even if the physical brick-and-mortar subway system, and its graph representation, would not. We discuss HONs in the next subsection and refer the interested reader to [24, 174, 69, 59, 118, 158] for further details.

Further variations

We note the above variations on the three main formalisms discussed are only the beginnings of possible ways to extend these representations. Depending on the complex system and questions at hand, certainly one may combine the variations described above to make, for example, an edge-weighted dynamic network [106], a weighted multilayer network [130], a multi-order network that combines multiple HONs [182], or another combination that provides an effective representation. One may also study systems of weighted nodes instead of weighted edges [192, 140], as well as representations where each node has some kind of internal structure [48, 71] or possible action [8] . Any of the formalisms above could also lend itself to studying the intricacies of coupled dynamical systems such as coupled oscillators [155, 147] or interacting threshold-linear models [136]. Indeed when including variations on the three formalisms covered in this review, we find we can encode an impressive range of complex system types and properties.

Other Formalisms

We recognize that many other formalisms intended for complex systems exist and that those we specifically mention in this review constitute only a small subset of the possibilities. Other possible formalisms include *graphons*, which describe limits of sequences of graphs and can be used to estimate large, noisy systems [33], *metapopulation models* which classically describe global behavior of many local species populations [120, 204, 93] and can be adapted to networks [48], random sequences of sets [25], and *sheaves* which can handle added information on each node in a network and have previously been used to frame the network coding problem [84] and find consensus in sensor networks [52].

3.5 Encoding system dependencies

As we discuss above, the formalism used to encode our data should be carefully chosen to respect any prominent properties of the system, and specifically the dependencies found therein. In this subsection we discuss the subtleties of choosing an appropriate formalism, and then review the common practices that researchers use to encode subset, spatial, and temporal dependencies using the formalisms we have introduced.

Once we have chosen which dependencies to model, it is important to carefully determine when two or more units in our system are related to each another – i.e. to define the relations in our model (Fig. 7.) Depending on the exact definition of the relations, the resulting representation may or may not exhibit the desired properties, or it may even exhibit properties not found in the actual system, but coming from externalities from the data, as discussed in Section 2.4.

For example, consider recording brain activity from an individual as they progress through different tasks (reading, watching a video, resting, etc.). Different tasks require the activation of distinct sets of brain regions. How do we define relations between brain regions? As depicted in Fig. 7, we could define k nodes to be related if a task requires all k nodes to be active. Alternatively, we could define a relation between k nodes if the k nodes were found to co-activate during a task. Finally we could call two nodes related if they have a high enough measure of pairwise similarity, perhaps assessed by correlation or mutual information. Depending on our chosen definition of node relations, our resulting representation either will or will not encode a subset dependency. In this example, only the definition of node co-firing exhibits a subset dependency, which we could capture in a simplicial complex representation. Now consider a city bus system fragment including stations, roads, and bus lines (Fig. 7, bottom). First, we could define a relation between k nodes (stations) as the sets of stations along an entire bus route. That is, k stations are related if they together form a whole bus route. This definition would propagate no subset or temporal dependencies to the representation. Second, we could instead call k nodes related if they share at least one bus line. Consequently we now have a subset dependency that must be captured by our choice of representation. Third, we might define two bus stations as related if they are subsequent stops along a route. This third, inherently pairwise, definition of relation could be represented with a graph. Note that none of these three definitions encode the temporal dependency, which may or may not be present in the available data. For example, if we had access to, not only stations' locations, but also passenger trajectories within the system, we could encode the temporal dependencies using HONs.

The above examples, and those in reference [195], illustrate the fact that one must carefully choose relations to effectively encode dependencies, or, equivalently, that whether or not a given representation exhibits a dependency is a (sometimes subtle) question of semantics. This is to say, the modeling choices concerning relations, representations, and dependencies are highly, and unavoidably, interdependent on one another. We must be aware of what dependencies exist in the system, which of those are encoded or neglected in the representation, and which come from external sources. In the scientific community, these difficult choices are usually made following the common practices that we delineate next.

Encoding subset dependencies

If a system exhibits subset dependency, it is common practice to use either simplicial complexes or hypergraphs to represent it. In the case when *any* subset of a set of related units are also related, then an appropriate formalism is the simplicial complex, since this formalism has the downward inclusion property (see Section 2.1). In the terms used in Section 2.1, the predicate P is true for any subset of an existing relation. If instead only *some* subsets of related units are related, then one could argue that a hypergraph is the appropriate formalism to use, since it allows for great freedom in encoding relations among subsets of related units. Equivalently, a particular subset dependency gives a particular choice of the predicate P, which in turn induces a particular hypergraph. Recall that the important difference between hypergraphs and simplicial complexes is the notion of a subedge. Drawing from Remark 3.5 of [195], if a 1-simplex $\{a, b\}$ and two 2-simplices $\{a, b, c\}$ and $\{a, b, d\}$ exist, then by definition $\{a, b\}$ is a sub-relation (formally called a *face*) of both $\{a, b, c\}$ and $\{a, b, d\}$. However, if instead we had hyperedges $\{a, b\}$, $\{a, b, c\}$, and $\{a, b, d\}$ in a hypergraph, we cannot say if $\{a, b\}$ is a sub-relation (sub-edge) of $\{a, b, c\}$, $\{a, b, d\}$, both, or neither. This connection or lack thereof between relations and sub-relations crucially affects interpretation of the system representation.

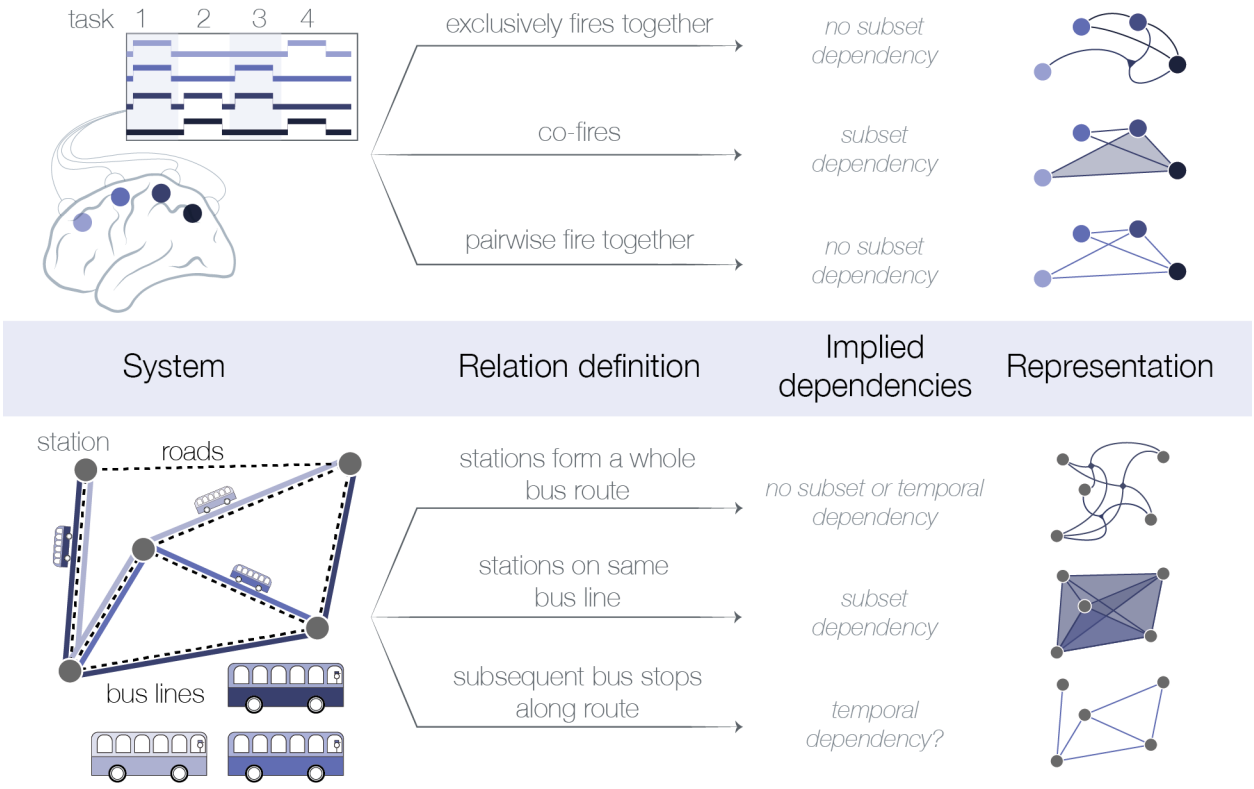

Figure 7: **Native dependencies captured by the definition of relation.** *(Top)* We might record the on/off activity of four brain regions in each of four tasks (left). Depending on the definition of relation chosen (middle), we may or may not record a dependency in a representation (right). *(Bottom)* Given five bus stations placed along a set of roads (dashed lines), we observe three bus lines that connect the stations (left). Depending on the definition of relation chosen (middle) we might include a subset or temporal dependency, which we would want to capture in our representation of the system (right).

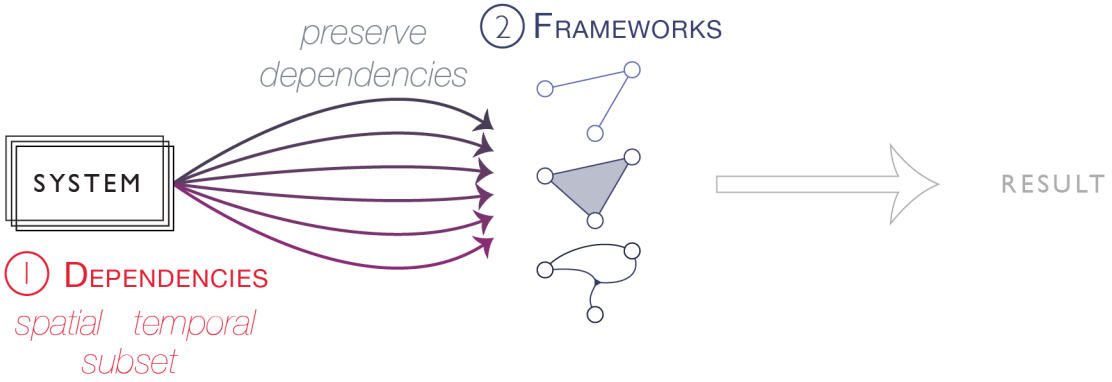

Figure 8: **Choosing a formalism marks the second step of our analysis pipeline.** Formalisms include graphs, simplicial complexes, and hypergraphs. We argue that the choice of formalism should be made in order to most faithfully capture system dependencies.

Encoding temporal dependencies

As discussed in Section 3.4, one way to encode temporal dependencies uses the idea of Higher Order Networks (HONs). Recalling our previous description, the HON begins with a set of walks, and from the patterns found therein creates a graph in which nodes correspond to ordered sets of units in the original system, and edges connect nodes based on temporal dependence. In this way, the HON takes the temporal dependency (for example, paths from A to B always lead to C), and encodes it in a special kind of node, derived from the original units of the system. At the time of writing, HONs have been defined for paths on graphs. It is still an open question how to extend the HON formalism to simplicial complexes or hypergraphs so that the resulting representation could exhibit both temporal dependencies and arbitrary subset (or spatial) dependencies.

Encoding spatial dependencies

Possibly the most straight-forward method to encode spatial dependencies constructs a weighted graph in which the edge weights in some way represent how close or far nodes lie from each other. However, we highlight the fact that edge weights are a popular mechanism to encode other kinds of information as well, and once we encode one piece of information within the edge weight we cannot then use edge weights to also encode spatial dependencies. For example, if we build a graph for the transportation system of a city, we may want to encode both traffic flow and road length as properties of the relations. Usually, both types of information are encoded using edge weights, so we are left with three alternatives. The first is to choose an edge weight that aggregates both types of information. The second is to use a multilayer network (see 3.4) in which each layer has weighted edges reflecting a single type of relation [44]. The third is to create a more holistic representation that efficiently combines the spatial information, traffic flow, and road length while also including any interactions between edge types. This challenge is yet another example highlighting that data availability, system dependencies, and choice of representation are not independent of one another.

If the only challenge to the study of complex systems were the choice of representation, then our discussion would be near complete. However, real-world systems usually have at least two or more dependencies, including those we do not discuss in this paper. For example, the subway network (Fig. 2.2) contains both temporal dependencies (evidenced in passengers' routes) and spatial dependencies (the routes taken are usually constrained by geographical proximity); while the co-author system (discussed further in Section 6)

could be further constrained by both temporal and subset dependencies. Moving forward (Fig. 8), we will need to develop novel methods for systematically representing and encoding complex systems with multiple dependencies.

4 Mathematical relationships between formalisms

At this point it may seem that the choice of representation wholly restricts the perspective and possible analyses on the data. For example if we encode the data as a directed hypergraph, we can only perform analyses using hypergraph methods. However as each of these base formalisms record relations between nodes, perhaps we could utilize the underlying mathematical relationships between each of these formalisms to gain additional insights. In this section we will explore the formal mathematical relationships between graphs, simplicial complexes, and hypergraphs, and then we will discuss the assumptions needed or information lost as we move from one to another.

From hypergraph to simplicial complex: Forgetting independent relations

First let us imagine that from our data we have constructed a hypergraph H. If we would like to create a simplicial complex K_H from H, we might first map the nodes of H to nodes of K_H, before dealing with the hyperedges. Recall that in a simplicial complex we have simplices that connect multiple nodes, but we also have the downward closure restriction that if we have a simplex r, then any $r' \subseteq r$ must also be a simplex. So then to form K_H we could take any hyperedge connecting $k+1$ nodes and form from it a k-simplex (Fig. 9 top left), thereby forcing the downward closure of the hyperedge relation so that the system representation can abide by simplicial complex rules. Additionally note that if we have a hyperedge a on nodes $\{v_0, \ldots, v_k\}$ as well as a hyperedge b on a subset of these nodes, the simplicial complex will view b as redundant information, since by definition every subset of nodes in a will be connected by simplices. In this way, we say that the simplicial complex "forgets" the existence of b as a relation observed independently of all other relations (specifically, observed independently from the relation a). We can also see this forgetting notion in the matrix representation of the structure itself: from a hyperedge incidence matrix we only need to keep the maximal hyperedge rows in order to build the corresponding simplicial complex incidence matrix. Additionally, K_H will also lose information regarding the total number of relations in which a node is involved, since many of those original hyperedge relations may be a subset of another hyperedge relation. On the other hand, this procedure allows us to access methods that are available for simplicial complexes but not for hypergraphs (discussed more in Section 5). Overall, in the hypergraph each hyperedge between a set of nodes arises independently, so that having additional hyperedges (or the lack thereof) between subsets of nodes within a larger hyperedge indeed supplies more information than the one largest hyperedge. In contrast, we can define a simplicial complex by its largest simplices (formally called maximal simplices) alone.

From simplicial complex to graph: Forgetting polyadic relations

Next let us assume that we are given a simplicial complex K, and that from K we wish to construct a graph G_K that still represents our data. This transition is more straightforward, as we can take all of the 1-simplices of K to be edges of the graph G_K. Said another way, if two nodes participate in the same k-simplex in K, then we draw an edge between these two nodes in G_K (Fig. 9, top right). By performing this transition from simplicial complex to graph, we are now forgetting polyadic relations between nodes. For example, in a simplicial complex we may have three nodes connected by three 1-simplices, or connected by three 1-simplices and a 2-simplex; in a graph, by contrast, we can only show these three nodes as being all-to-all connected by edges thus eliminating our ability to distinguish between the two cases. One can also move from a hypergraph to a graph by drawing an edge between two nodes only if the two nodes were connected by a hyperedge. The resulting graph recovered from this process will be the same as the graph obtained by moving from a hypergraph to a simplicial complex to a graph following the described protocol.

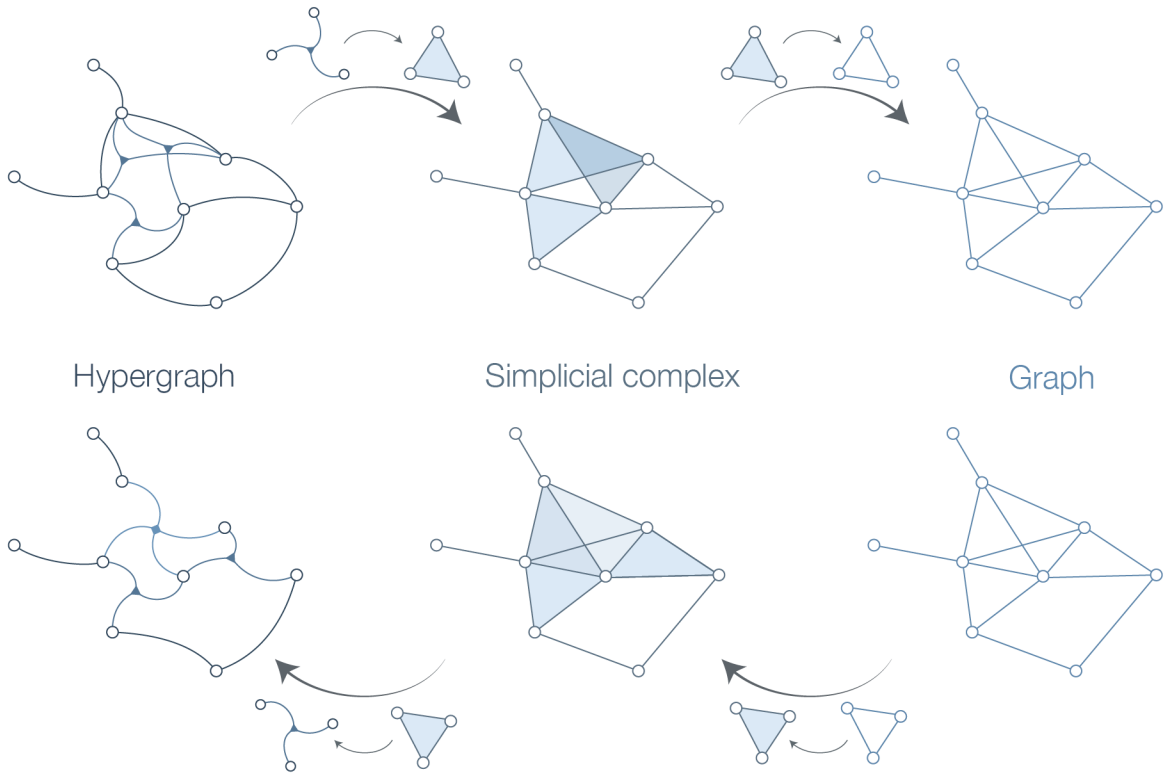

Figure 9: **Transitioning between formalisms requires added assumptions or engenders forgetting information.** *(Top, left)* The original example system as a hypergraph. *(Top, middle)* When we send hyperedges to simplices, we create a simplicial complex, and *(Top, right)* by keeping all edges we form a graph. Going in the other direction, we begin with the same graph *(bottom right)*, fill in all cliques as simplices to obtain a simplicial complex *(bottom middle)*, and send maximal simplices to hyperedges to form a hypergraph *(bottom left)*. Note that the hypergraphs on the top left and bottom left differ from one another.

From graph to simplicial complex: Assuming polyadic relations

What happens if we instead move in the other direction? What are the assumptions necessary to take a graph such as the graph shown in Fig. 9, right, and construct from it a simplicial complex or a hypergraph? First, let us begin with a graph G and construct a simplicial complex. If we make the assumption that all nodes involved in a $(k+1)$-clique of G are related, then we can construct a simplicial complex K_G by filling in each $(k+1)$-clique with a k-simplex. This particular construction is called the *clique complex* [104] or the *flag complex* [105], and is often denoted by $X(G)$ (Fig. 9, bottom middle). We reemphasize that for this construction, it is necessary to assume that all nodes within a clique are all together related as a single functional unit. Importantly this clique-to-simplex assumption may not be appropriate for all systems. One example arises from social conversations in which three people may converse only in pairs and never together as a three-person group.

From simplicial complex to hypergraph: Assuming that only maximal simplices are independent

As we consider moving from simplicial complex K to hypergraph H_K we are faced with a few options. First, since a simplex by definition implies that all subsets of nodes within a simplex are also related, then we could take every simplex and form from it hyperedges between all subsets of nodes within the simplex. In constructing the hypergraph in this way, we carry through the downward closure restriction. Alternatively, we could perform a conversion more akin to the inverse of the hypergraph-to-simplicial-complex conversion discussed above by creating a hyperedge only for each maximal simplex of K (Fig. 9 bottom left). Assigning hyperedges only for maximal simplicies can be seen as a conservative approach; that is, we can uniquely define a simplicial complex using its maximal simplices so that in forming the new hypergraph, we are assuming the fewest number of hyperedges necessary to preserve only the polyadic relations already known to be independent.

From hypergraph to graph, and from graph to hypergraph

Perhaps most importantly, note that from a graph we can move to a simplicial complex, then to a hypergraph, then back to a simplicial complex, and finally back to a graph following the translations discussed above. In this process, we will recover the original graph with which we began. However, the opposite is not the case. As depicted in Fig. 9, we can begin with the hypergraph on the top left, move through the simplicial complex, to the graph, and then move back along the bottom row from right to left and we will in fact recover a very different hypergraph than the one from which we began. This exercise emphasizes the information lost or forgotten in moving down the formalism ladder. Specifically, since each hyperedge may arise independently of all others (most notably independently of any hyperedge that is a superset), we not only lose information when moving towards a graph but also cannot recover this information when moving back up from a graph to a hypergraph.

We note that the above protocols of moving from one formalism to another do not encompass all possibilities. One could define a simplicial complex from a graph by simply keeping all edges as the 1-skeleton and having no larger simplices. Or perhaps one might form a weighted graph from a hypergraph by assigning edge weights as some function of the hypergraph structure [141, 43, 91]. Though we here discussed moving between formalisms as the third step in the pipeline (Fig. 10), moving from one formalism to another *after* the initial encoding of data into a formal representation should be performed only with extreme care, as any translation requires adding assumptions or forgetting relations or independencies.

5 Methods suitable for each representation

Now that we have exerted the effort necessary to properly represent our data as a graph, simplicial complex, or hypergraph, how do we analyze the resulting structure? In this section, we will describe methods

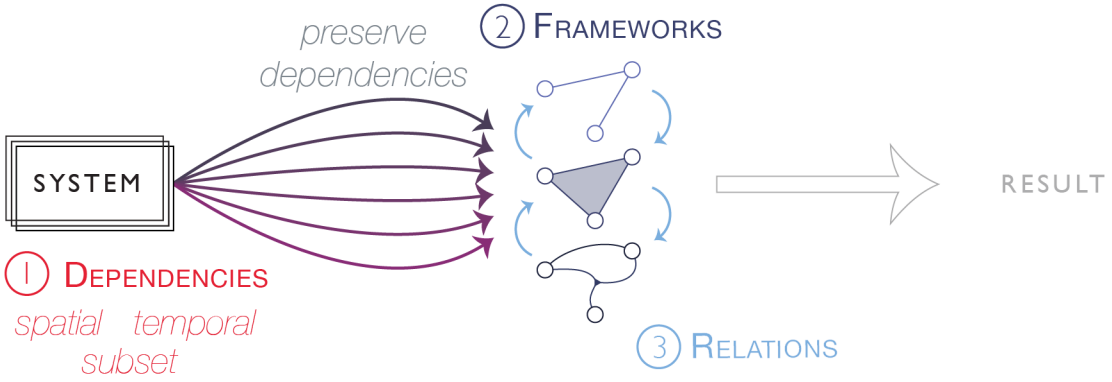

Figure 10: **Mathematical relations between formalisms add an optional step to the pipeline, only to be used with caution.** After an appropriate formalism has been chosen for system representation, one can make use of mathematical similarities between formalisms to view their representation from the lens of a different formalism. Importantly, we note that moving from hypergraphs to simplicial complexes to graphs can result in a loss of information, while the reverse direction can require one to make assumptions about the system.

Figure 11: **Formalisms provide different perspectives on the neighborhood of a node.** *(Left)* The colored node has four direct neighbors and participates in two triangles. *(Middle)* The colored node has four direct neighbors and participates in one 2-simplex. *(Right)* The colored node has two neighbors connected to itself exclusively, and two neighbors accessible through a larger hyperedge.

and statistics that can be used to evaluate precisely how each of the three base formalisms offer unique perspectives on the system under study. We recognize that many such methods and statistics exist, but for clarity we will focus on a few techniques that help us identify similarities among and differences between representations.

Before we begin, we briefly provide a note of caution. The fact that a method of interest might currently intake only one particular formalism does not justify the use of that formalism in representing our data. To further illustrate the point, if we intend to understand the spread of a disease by way of calculating the epidemic threshold [41], we would find that most existing methods that calculate the epidemic threshold do so from a graph representation. The theory of disease spread on hypergraphs and simplicial complexes currently is a nascent area of research [99, 101], so one might not find a definition of the epidemic threshold that uses either of these polyadic formalisms in the literature and is appropriate for the system at hand. Nevertheless, the absence of this particular notion for polyadic formalisms does not imply that we are justified in using a graph formalism to represent the system. Generally, a result is unlikely to offer fruitful insight into a system if the calculation was performed on a representation that itself is ill-suited for the system.

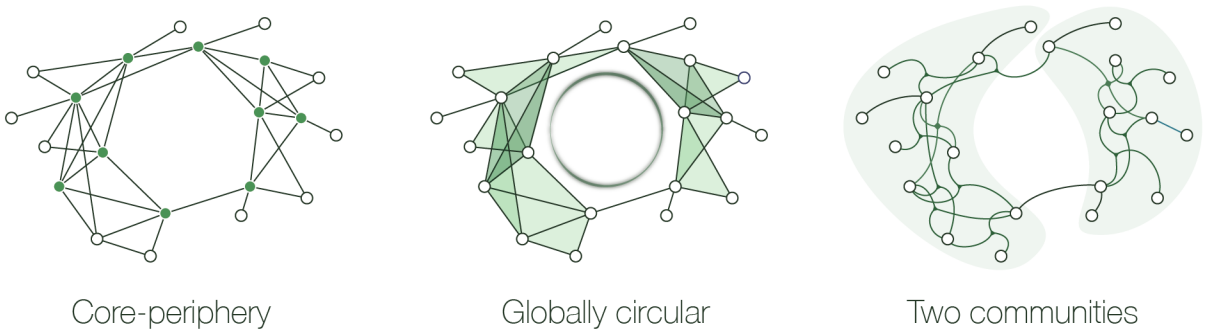

Figure 12: **The three formalisms and their corresponding downstream analyses can offer different perspectives on a complex system.** *(Left)* For this example system, a graph representation suggests a global core-periphery organization. *(Middle)* A simplicial complex representation of the same system appears to show a globally circular structure. *(Right)* The hypergraph representation of the same system hints at the presence of two communities.

5.1 Methods for graphs

As the most well-known of the three formalisms in data analysis, graphs have offered scientists interpretable and easily-computable tools for centuries. Thanks to this rich history of graph analysis, we can computationally study graphs at many different levels: the local node or node-neighborhood level, a meso-scale level to see larger patterns, and the global level to summarize the entire object. Though myriad statistics exist, for the sake of brevity we limit our discussion below and point the interested reader to [142] to learn more.

At the most basic level, the number of edges incident to a node v_i is called the node *degree* and is denoted k_i (Fig. 11, left). The distribution of degrees can constrain the graph's large-scale organization, for example tracking the emergence of a giant connected component [134]. At the neighborhood level, we can investigate statistics describing the connectivity between a node's neighbors. A common example is the *clustering coefficient* c_i of a node v_i. Formally the clustering coefficient is

$$c_i = \frac{2\mu_i}{k_i(k_i - 1)}\,, \tag{1}$$

where μ_i is the number of edges between neighbors of v_i. The numerator counts the number of triangles in which v_i participates and the denominator normalizes by the number of triangles that could possibly form around v_i. Broadly, the degree and clustering coefficient are examples of a much broader class of statistics proposed for the description of local and neighborhood structure in graphs.

Complementing such descriptions, other statistics have been defined to measure the nature of paths in the graph and markers of meso-scale structure. For example, the average path length, various types of centrality [79, 22], notions of modularity [90, 114, 151], and the property of small-worldness have proven useful in the study of a wide variety of systems from the human brain [36, 17] to granular materials [156]. One particular method that, at the time of writing, we found to be unique to the graph formalism, is that of core-periphery structure. A graph with core-periphery structure contains a dense group of nodes connected to each other called the *core*, and a second group of nodes called the *periphery* that mostly connect to the core rather than to other nodes in the periphery [32, 172] (Fig. 12, left). For a description of other network statistics, we refer the interested reader to prior literature [142, 175]. Additionally, we note that in real world systems, the values of many of these network statistics are correlated with one another over instances in a graph ensemble, and these patterns of shared variance can be used to distinguish between types of systems [50, 146].

5.2 Methods for simplicial complexes

Simplicial complexes entered the data analysis scene more recently than graphs. Yet, we can still use a simplicial complex to investigate multiple levels of system architecture with intuitive methods and statistics. As before, we keep this section brief by focusing only on a few basic statistics and then one method that is unique to simplicial complexes.

We might first seek to extrapolate basic graph definitions to simplicial complexes. If we view a graph as the 1-skeleton of a simplicial complex, then the graph degree of node v_i corresponds to the number of 1-simplices in which v_i participates. By extending this idea, we can understand the neighborhood of a node (Fig. 11, middle) by defining the *simplex participation* of node v_i as the vector $P(v_i)$ in which the k^{th} element is the number of $(k-1)$-simplices in which v_i participates. One could also record the vector of simplices in which the node participates (called the upper degree in [185]), or the number of simplicies not contained in any larger simplices, i.e. the maximal simplices, in which the node participates [191]. Notably, one can also define the degree of a simplex of any size, instead of only defining the degree of a node (which are after all, zero-dimensional simplices) [185]. Similarly, we might ask whether and how the clustering coefficient could be extended to the simplicial complex formalism. Depending on the precise properties that one intends to capture, one could use a ratio of simplices from dimensions k and $(k-1)$ to formalize the notion of a clustering coefficient. However, in the simplicial complex formalism each simplex can be considered as a fundamental building block, so it makes sense to also define a clustering coefficient for an arbitrary k-simplex as in [125]. In a complementary effort, the notion of centrality has recently been extended from the graph formalism to the simplicial complex formalism [73].

In addition to extending graph measures to simplicial complexes, we can also harness the theory of algebraic topology to uncover more complicated motifs within the system. The downward closure requirement within the simplicial complex definition gives us the ability to accurately identify which simplices are involved in higher dimensional simplices. Consequently we can then detect where a dearth of simplices leaves topological voids in the complex (Fig. 12, middle). Detecting topological voids is the work of *homology*, and as homology relies on well-defined mappings from larger to smaller simplices, this method is best suited for the formalism of simplicial complexes.[3]

Simplicial complexes can also be "reversed" in a way that can be useful in understanding the structure of grouped nodes, while preserving the topological organization of the system. Consider constructing, for example, a simplicial complex in which nodes represent regions of the zebrafish brain, and simplices represent co-activity during a task. We could encode the complex as a $\#simplices \times \#nodes$ binary matrix sometimes also called a *concurrence* matrix or *incidence* matrix [85, 64]. In the top left of Fig. 13 we show a small example concurrence matrix of five nodes ($1, 2, 3, 4$, and 5) connected through four relations (a, b, c, or d). We create a simplicial complex (Fig. 13, top right) by drawing maximal simplices between nodes that share a relation in the concurrence matrix. For the zebrafish example, the simplicial complex could contain relatively few simplices but orders of magnitude more nodes (depending on data availability of course), making calculations cumbersome. As an alternative, we could "reverse" the structure by constructing the *Dowker dual* [64]. Here, the role of nodes is swapped with the role of relations [64]. In the zebrafish example, we would form a node for each co-activity relation, and then connect two nodes by simplices if they share a participating region in the zebrafish brain. In Fig. 13 we transpose the concurrence matrix to swap the role of nodes and relations, and then show how we again create a simplicial complex now called the Dowker dual. This new complex will have the same number of nodes as the original complex had maximal simplices, so that if the number of relations was small with respect to the number of nodes in the original complex, studying the Dowker dual will be more computationally tractable.[4] Importantly, studying the Dowker dual preserves specific topological structure within the system [64]: the homology groups of a simplicial complex and its Dowker dual are isomorphic. Thus, the Dowker dual can be an incredibly efficient representation, assuming that we still respect the scientific question at hand.

[3]Sometimes the words "structural" and "topological" are used interchangeably. In this work, we use the adjective "topological" to modify nouns relating to the theory of algebraic topology. We use "structural" to generally refer to the patterns formed by the units and relations of a system.

[4]This construction is akin to the *line graph* construction in graph theory [94].

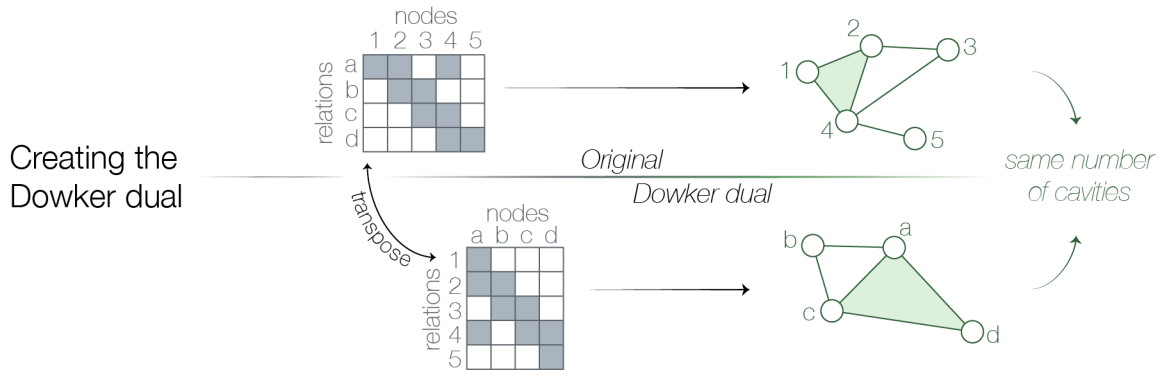

Figure 13: **Constructing the Dowker dual of a simplicial complex.** Given a concurrence matrix denoting which nodes connect via relations (top left), we can create a simplicial complex with each simplex defined by a relation (a, b, c, d). Alternatively, we could transpose the matrix so that now we have four nodes (a, b, c, d) and five relations (bottom left). From this transposed concurrence matrix we can then create a simplicial complex whose simplices are defined by relations in the new concurrence matrix (bottom right). This simplicial complex is called the Dowker dual of the original simplicial complex, and will have the same number of topological cavities.

5.3 Methods for hypergraphs

Like simplicial complexes, hypergraphs have gained popularity only recently, as many fields have begun to realize the importance of encoding polyadic relations in systems. The hypergraph formalism is a natural extension of the graph formalism, so unsurprisingly many (though certainly not all) computational methods for hypergraphs are extensions of computational methods for graphs. As before, we will highlight basic statistics here as well as a method unique to hypergraphs.

The formalism of hypergraphs is also complemented with a set of descriptive statistics. Importantly, recall that each hyperedge arises independently since we have no rules relating hyperedges to each other, and consequently computations on hypergraphs must be interpreted differently from related computations on simplicial complexes. Starting simply, we can first extend the concept of degree to hypergraphs. In a hypergraph, the degree of a node $d_H(v_i)$ is the number of hyperedges containing v_i, sometimes called the *hyperdegree*. Since hyperedges can connect any number of nodes, we also define the *hyperedge cardinality*, also called the *hyperedge degree*, as the number of nodes contained by the hyperedge. Importantly, note that in the definition of node degree, we do not stratify by hyperedge cardinality as the appearance of a large hyperedge gives no information about the existence of smaller hyperedges. Instead, a large hyperedge is simply another relation that contains our node of interest.

In order to understand a node's neighborhood in a hypergraph (Fig. 11, right), next we move to a definition of the hypergraph clustering coefficient (see [111, 72, 157, 80] for others). Recall the graph clustering coefficient measures connectivity of a node's neighbors via connections that *do not* include the node of interest. If we examine, for example, node v_i and its neighbors, some neighbors will be connected via hyperedges that do or do not include v_i. Intuitively, node v_i should have high clustering if its neighbors connect via hyperedges that do not contain v_i. The *extra overlap* $EO(v_i)$ of a node v_i helps us to quantify this idea; formally, the extra overlap of two hyperedges e_j, e_k is defined as

$$EO(e_j, e_k) = \frac{|N(D_{j,k}) \cap D_{k,j}| + |D_{j,k} \cap N(D_{k,j})|}{|D_{j,k}| + |D_{k,j}|}, \tag{2}$$

where $D_{j,k} = e_j \setminus e_k$, and $N(U)$ is the set of all nodes that are neighbors of any node within the set U. Then intuitively the extra overlap between two hyperedges counts the number of nodes connected by outside hyperedges, and we normalize by the size of the two hyperedges under consideration. Note that if we have

only hyperedges of cardinality 2, then the extra overlap over two edges involved in a triangle is 1. Finally, the hypergraph *clustering coefficient* $C_H(v_i)$ of a node v_i is

$$C_H(v_i) = \begin{cases} \binom{|M(v_i)|}{2}^{-1} \sum\limits_{e_j, e_k \in M(v_i)} EO(e_j, e_k) \text{ if } d_H(v_i) > 1 \\ 0 \text{ if } d_H(v_i) = 1 \end{cases}$$

where $M(v_i)$ is the collection of hyperedges that include v_i [228]. This definition for the hypergraph formalism is thus similar in spirit to the definition of a clustering coefficient for a graph. Indeed, the former is equivalent to the latter when all hyperedges have cardinality 2.

We note that the hypergraph also has the ability to uniquely represent the absent substructures of a system. Much like identifying repeated structural patterns (or *motifs*) in a graph, a hypergraph allows us in principle to identify repeated patterns of *absent* hyperedges. We may have a case where, for example, pairwise hyperedges exist between four nodes that also connect via a 4-hyperedge, but no 3-hyperedges exist. An interesting research question for such a representation is to ask why we observe a lack of three-node relations but an abundance of 2-node relations within every 4-node relation. Note that neither graphs nor simplicial complexes allow for this line of questioning due to the lack of polyadic relations or the requirement of downward inclusion, respectively. A detailed investigation of these absent substructures is outside of the scope of this paper; yet, we can take a step in that direction by defining the following statistic, which we call the *fill coefficient* of a hyperedge h, as

$$f(h) = \frac{|g \in \mathscr{E} : g \subsetneq h \text{ and } |g| > 1|}{2^{|h|} - 2 - |h|}, \tag{3}$$

where $\mathscr{E}$ is the set of hyperedges, and $|\cdot|$ is the cardinality of a hyperedge. The fill coefficient intuitively describes the fraction of smaller hyperedges that exist within hyperedge h, taking into account the hyperedge cardinalities.

5.4 Methods and dependencies

Before closing this section, we note that both in choosing analyses and in interpreting results, we need to keep in mind the dependencies within the system. For example, after creating a simplicial complex from our data, how do we interpret its clustering coefficient? Or what does the diameter of a system mean when we have hyperedges of different cardinalities linking nodes instead of (dyadic) edges? How do communities found from a simplicial complex [30] with subset dependencies differ from communities found within a hypergraph [107] without such dependencies? Can we intertwine different sorts of system dependencies to understand their impact on function? Examples of such intertwining methods include (i) Rentian scaling, which formalizes the interaction between structure and geography [45, 16, 154], and (ii) modularity maximization with spatial null models [76, 28]. Careful consideration of the above questions can only lead to better motivated, more interpretable, and insightful results.

To summarize, we see that each base formalism offers a particular perspective on the data it encodes. As we show in Fig. 12, the choice of formalism can influence how we interpret the complex system structure of the same dataset. The graph representation could suggest a core-periphery structure; the simplicial complex representation lets us see that globally the system organizes around one circle; and the hypergraph representation highlights the existence of two communities. We close this section by emphasizing the importance of formalism choice for proper representation of the data, and the appreciation that each analysis performed or pipeline chosen offers a different perspective on the underlying complex system (Figure 14).

6 Examples

Putting it all together, in this section we will discuss examples of a system, its dependencies, and how we might represent the system using the formalisms described above. We will then explore possible anal-

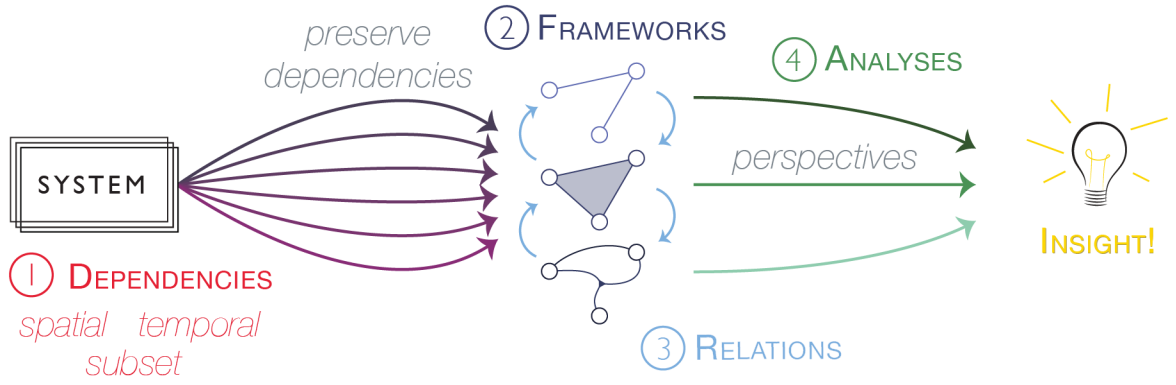

Figure 14: **Updated analysis pipeline includes consideration of which computational methods to perform on the chosen representation, and what distinct or complementary perspectives these methods offer.** The last step in our pipeline involves computationally analyzing the system representation. We note that each analysis provides its own perspective on the system representation. We recommend performing steps 1-4 with careful consideration in order to gain real insight into the system.

yses on each representation and compare results. Importantly, we will see that the subtle differences in representations and their definitions can lead to inconsistent results and conflicting interpretations.

6.1 Co-authorship

Consider the system made up of scientific researchers who interact to write scientific papers (for example, [46]). What kind of dependencies govern the relations in this system? How should we encode this system formally? In the following paragraphs we analyze a fragment of this system following the workflow of Fig. 14. We begin with a toy example (Fig. 15) and later perform similar analyses on a real dataset (Fig. 16).

Dependencies First, we may expect to find spatial dependencies in this system, as the country of origin or university affiliation of a researcher may dictate which of their colleagues are willing or able to collaborate. Second, we may also expect temporal dependencies, as a researcher's past collaborators may also influence any future collaborations. Third, whether this system exhibits subset dependencies depends upon which precise fragment we are interested in studying. If we focus solely on authors and consider two or more authors as related whenever they have worked together at some point, then it is necessarily the case that whenever a group of three authors have co-authored a paper, then any two of them have co-authored a paper, and therefore a polyadic relation always implies all smaller polyadic sub-relations, including all dyadic sub-relations. On the other hand, we may choose to focus on both researchers and scientific papers, in which case, we may want to think about one scientific paper as determining a single relation. In this scenario, the fact that three researchers are involved in a relation (because they have authored a paper all together) does not imply that two of them have authored (another, separate) paper together. We will keep these dependencies in mind as we move through the later analysis steps.

Externalities: data availability In a vacuum, we may expect to see all of the aforementioned dependencies in this system. However, the data available may be biased in such a way that, for example, researchers working (and papers produced) in one particular country are over-represented. In this case, the data available may not adequately record the spatial dependencies involving, for example, researchers who travel frequently between two different countries. Moreover, if the dataset is further biased to include solely researchers that

work in one particular institution, then it is possible that no spatial dependencies are recorded at all, since researchers in one institution may all work with one another with the same likelihood; that is, the location of one existing collaboration offers no new information about the likelihood of another collaboration.

Externalities: research question Let us introduce a toy example to accompany our discussion. Consider a co-authorship dataset including four authors a_1, a_2, a_3, and a_4 who have written four papers p_1, p_2, p_3, and p_4 (Fig 15, top). The three papers were authored as follows: paper p_1 was authored by $\{a_1, a_2\}$, paper p_2 by $\{a_2, a_4\}$, paper p_3 by $\{a_1, a_2, a_3\}$, and paper p_4 by $\{a_3, a_4\}$. This toy example illustrates that whether the chosen representation reflects the dependencies inherent to the system is sometimes a subtle question of semantics. For example, if the relations in our representation are defined using the first question ("has this pair of authors worked together on at least one paper?"), then the set of relations will necessarily exhibit a subset dependency, regardless of whether or not the data available records a subset dependency found in the real system. Said another way, one must be aware of which dependencies reflected in our representations come from the system, from how the data was collected, or from the representation constructed. In this example it is the question at hand, the intricacies of the system under study, and the data available that all together guide the choice of representation of these data.

Representations If we take the information from Fig 15, top left, and construct the classic co-author network in which an edge exists between two authors if they have appeared as co-authors on a paper, then we recover the graph shown in Fig. 15, top right. In particular, note that as we construct the co-authorship graph, we ask the following question exactly once for each potential relation: "Has this *pair* of authors worked together on at least one paper"? Alternatively, we can consider polyadic relations between authors and ask "has this *set* of authors worked together on a paper?" This question naturally yields a simplicial complex (Fig. 15, middle right), in which nodes form a simplex if the corresponding authors are a subset of the authors of at least one paper. Finally, we imagine the author list of the paper is non-redundant so that one paper corresponds to exactly one relation. Said another way, we respect that without each and every author, the paper could not have been completed. If we take this point of view, we will instead construct a hypergraph by repeatedly asking "Has this set of authors exclusively (needing no other authors) written a paper together?" This approach retains the large group of three authors, but now clearly shows that, for example, authors a_1 and a_3 have not worked on a project as an exclusive group. Note that this information is not recoverable from either of the other representations.

Importantly, access to the full data about researchers and scientific papers would allow us to build any of the three formalisms discussed. However, if we only have information about co-authorship (which sets authors have worked with each other) rather than full knowledge of the data, we would only have been able to construct the graph version or the simplicial complex version, but not the hypergraph version[5].

Methods and Analyses Next we analyze the three different system representations. Of our co-author representation (graph, simplicial complex, or hypergraph), we might first ask a simple question about the involvement of a node (author) in paper writing in order to gauge the author's productivity. In the graph, we might use the node degree to recover this information, which would tell us that authors a_2 and a_3 participate in the same number of collaborations. Moving to the simplicial complex, we see by looking at node participation in maximal simplices that again authors a_2 and a_3 could be described as equivalently collaborative. However, in the hypergraph representation, if we look at node degree we see clearly that a_2 has participated in more collaborative projects than any other author, a conclusion that we are only able to draw from the hypergraph representation. A similar experiment comparing authors a_1 and a_4 shows that in this scenario, both the simplicial complex and hypergraph encodings view these authors as having different sizes of collaborative projects, while the graph structure does not. Specifically, the graph tells us that both a_1 and a_4 have worked with a_2 and a_3; the simplicial complex tells us that a_1 worked collectively with a_2 and a_3 whereas a_4 only worked individually with a_2 and a_3; and finally the hypergraph tells us that a_1

[5]One could build a hypergraph version, but it would not be an appropriate representation for this scenario because it does not respect the blatant subset dependency.

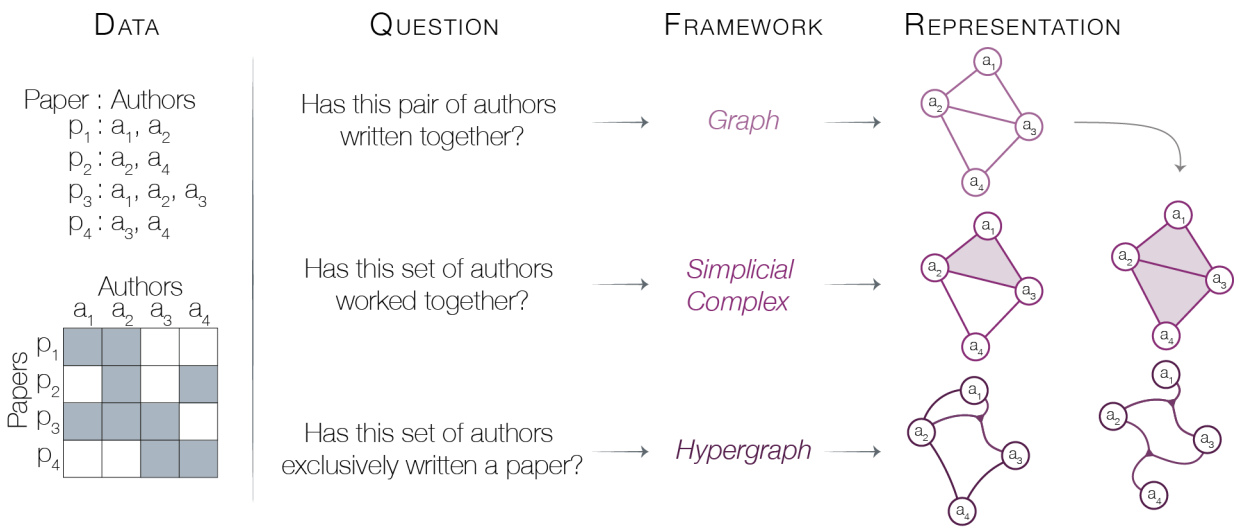

Figure 15: **Example of different perspectives offered by each formalism on a co-author dataset.** The data (far left) consists of a list of papers and their author list. Based on a question about what defines relations between authors, we build either a graph, simplicial complex, or hypergraph (right). If we started with the graph, we could also use the relations between formalisms to create a simplicial complex or hypergraph (far right), though this process can result in inaccurate representations of the original data.

had an individual project with a_2 as well as a team project that also included a_3 while a_4 only worked on two-person papers. These analyses illustrate how the subtle differences in the three discussed representations and associated downstream analyses can yield insights that may be at odds with one another.

Relationships between formalisms In this toy example we constructed each of the three representations directly from the data itself, with full knowledge of the raw data. However, we could also imagine that we are given the data already represented as one formalism and then try transforming our representation to another formalism. If we begin with one representation and translate to another formalism, will we recover the same information as if we had constructed the structure directly from the data? Here if we begin from a graph and move to a simplicial complex by attaching simplices to cliques (i.e. construct a *clique complex*), we recover a simplicial complex with two maximal simplices formed by a_1, a_2, a_3 and a_2, a_3, a_4 (Fig. 15, far right). Moving then from simplicial complex to hypergraph we would form a hypergraph with two hyperedges between a_1, a_2, a_3 and a_2, a_3, a_4; see Fig. 15, far bottom right. If we asked the same questions about author participation as we did above, then we would find that both pairs (a_2, a_3 and a_1, a_4) now seem to contribute in exactly the same way across representations. In studying complex systems we may receive only one representation of the system rather than the raw data, which can make switching to a different representation that perhaps better suits the planned analyses enticing. However, the present exercise underscores the importance of understanding the assumptions made by each formalism; care must be taken when moving between formalisms, not simply in recasting the mathematical language used, but also in remaining true to the original data.

Real dataset example We close this example by illustrating the above points in a real co-authorship dataset extracted from the DBLP computer science bibliography database [23]. This dataset consists of 3,700,681 scientific articles published between the years 2000 and 2016, as well as the list of authors of each article, for a total of 1,930,378 authors. Using this dataset, we build separately a graph, a simplicial complex, and a hypergraph directly from the data for each year contained in the dataset. In each representation, we measure the degree of each node, using the definitions in Section 5. Figure 16a contains a scatter plot showing

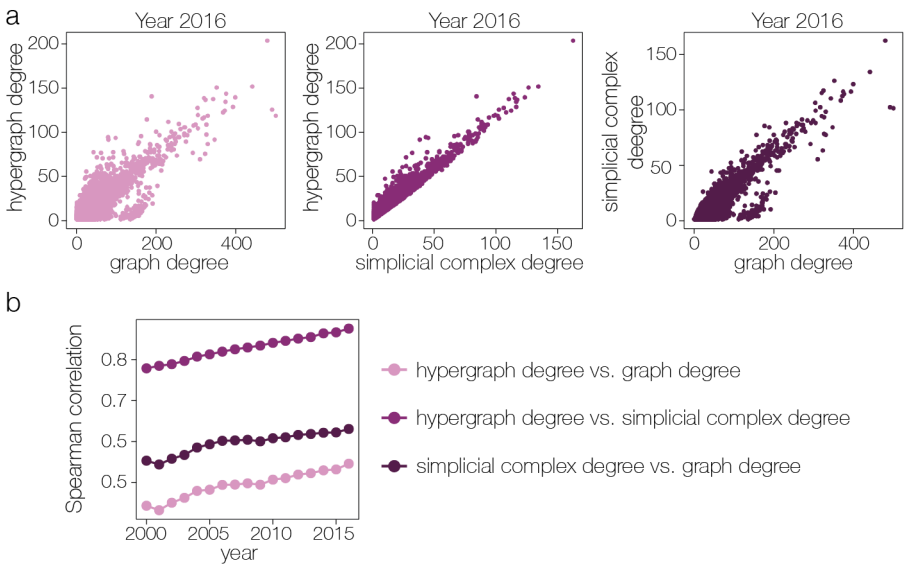

Figure 16: **Correlation among degree measurements in different representations of the same dataset.** *(a)* Comparing the degree calculated from the graph or hypergraph representation (left), from the simplicial complex or hypergraph representation (middle), and from the graph or simplicial complex representation (right) from the co-authorship dataset extracted from the DBLP computer science bibliography database in year 2016. Correlation is relatively high for nodes of large degree, but relatively low for nodes of small degree. *b* Spearman correlation coefficient calculated between node degrees from pairs of data representations in each year.

the degree of each node as measured in the different representations corresponding to year 2016. In this dataset, the degree in any representation is positively correlated to the degree in any other representation, though progressively less so as the degree of the node decreases. This result means that the different representations often agree more on which nodes have the largest degrees than on which nodes have small degree. This is important to keep in mind, especially in studies that make claims about the nodes of small degree, which often outnumber those with large degree.

To see how this correlation changes over time, for each year we calculate the Spearman rank correlation coefficient, which quantifies the similarity in node degree rankings between two representations. The Spearman rank correlation coefficient is equal to 1.0 when the rankings are equal, and -1.0 when the rankings are exactly reversed. Shown in Fig. 16b, we measure this coefficient for each year and each pair of representations. We observe the highest correlation between degrees calculated from the simplicial complex representation (number of maximal simplices) and degrees calculated from the hypergraph representation (number of hyperedges), which is likely due to the fact that these two representations both encode the polyadic relations in the dataset. This result suggests that relatively few papers authored by a subset of the authors of another paper were written. We also observe that the degrees calculated from the graph representations show a comparatively low correlation to the node degrees calculated from the other representations. In particular, the correlation between the graph and the hypergraph drops below 0.5 in some years, signaling a very different result when ranking nodes by graph degree or by hypergraph degree. Our observations imply that, in this dataset, we should be careful when making broad claims regarding the degree of nodes, especially those with few observed relations (i.e. small degrees), as each representation may yield different results that must be interpreted accordingly.

In summary, we have used this example to illustrate each step of the workflow from Figure 14, as well as to show that using different representations of the same dataset may yield measurements that are at odds

with each other, even in the simple case of measuring node degree and when the representations are created directly from the data.

6.2 Email communications

In our next example we again follow the the workflow of Fig. 14, but more succinctly. While in the previous example we discussed multiple types of dependencies, variations in data availability, and differing research questions, here we provide an example of a seemingly straightforward analysis on a dataset of emails.

We start by considering the following scenario. Suppose Ana works at a company and is tasked with improving communication and cohesiveness between teams in the workplace. Ana works at a big company that contains many teams in diverse areas, so she decides to prioritize her involvement by focusing on average team communication via an easily accessible medium such as email. Concretely, Ana wants to evaluate how well each team integrates with all members of the company, which translates to evaluating the average clustering coefficient of each team. For this purpose, Ana has collected all the internal email communications. She decides to operationalize her task as follows. First, if a set of at least 5 people have all received the same email at the same time, she will assume they must be working together as a team. Second, she decides to focus on emails with at most 25 participants, as emails with more than 25 participants are likely company-wide communications that do not involve a single team working together. Third, having identified a team, she will quantify the team's cohesiveness by averaging the clustering coefficient of each member in the team. Note that in this system we represent no spatial dependencies as email allows instant communication regardless of geographical location. Finally, Ana only cares about the teams and communication that have already occurred, and is not hoping to predict communication in the future, so for the presented analysis on aggregate communication she does not need to incorporate any temporal dependency that might exist within the system.

To follow up with her plan, Ana needs to choose a formalism with which to encode the data, as well as how to measure the clustering coefficient. If she chooses to encode the system as a graph where each employee is a node and each edge joins two nodes if they simultaneously received the same email, then she may use the clustering coefficient defined in Section 5.1. Alternatively, she may choose to build a hypergraph where each node is an employee and each hyperedge denotes a single email, and use the clustering coefficient defined in Section 5.3. A priori, one might expect that measuring the average clustering of a set of nodes in the graph is highly correlated to measuring the same quantity in the hypergraph. However, we will see that the subtle differences in definitions lead to varying results.

For this example, we use a dataset of email communications [23], containing $10,883$ emails among 148 employees of a company, from March 1999 to October 2002. Each email has a corresponding set of participants which includes the sender and all recipients. In Figure 17 we see the results of Ana's analysis on this dataset, using both a graph and a hypergraph representation of these data. Each marker represents an email with between n and 25 participants, for $n = 5, 6, 7, 8$, i.e. what Ana considers to be a team. Each email is located according to the average clustering coefficient of the email's participants, as measured in the graph (horizontal axis) and in the hypergraph (vertical axis). In the top left panel we can see that there is very little correlation between these two quantities for teams of at least 5 people (Spearman rank correlation coefficient $r = 0.1$, and associated p-value $p = 0.16$). These results show that ranking teams of employees using these two different clustering coefficients yields very different results. Consequently, if Ana wants to prioritize teams in order of how much she needs to intervene, in other words by ranking the teams according to average clustering coefficient, then the graph and hypergraph representations will recommend very different courses of action. Indeed, Ana would need to allocate her resources in entirely different ways depending on which representation she chose. This result does not change if Ana chooses to focus on teams of at least $6, 7$, or 8 people, as shown in Fig 17.

Next, Ana decides to distinguish between team cohesiveness with the company and intra-team cohesiveness. That is, do those teams that communicate well with everyone in the company also have robust within-team communication? A team that has excellent internal communication would have a high fill coefficient, which measures the fraction of possible smaller hyperedges that exist between the nodes of a hyperedge (see Section 5.3). In order to answer this question Ana calculates the fill coefficient of each team and adds

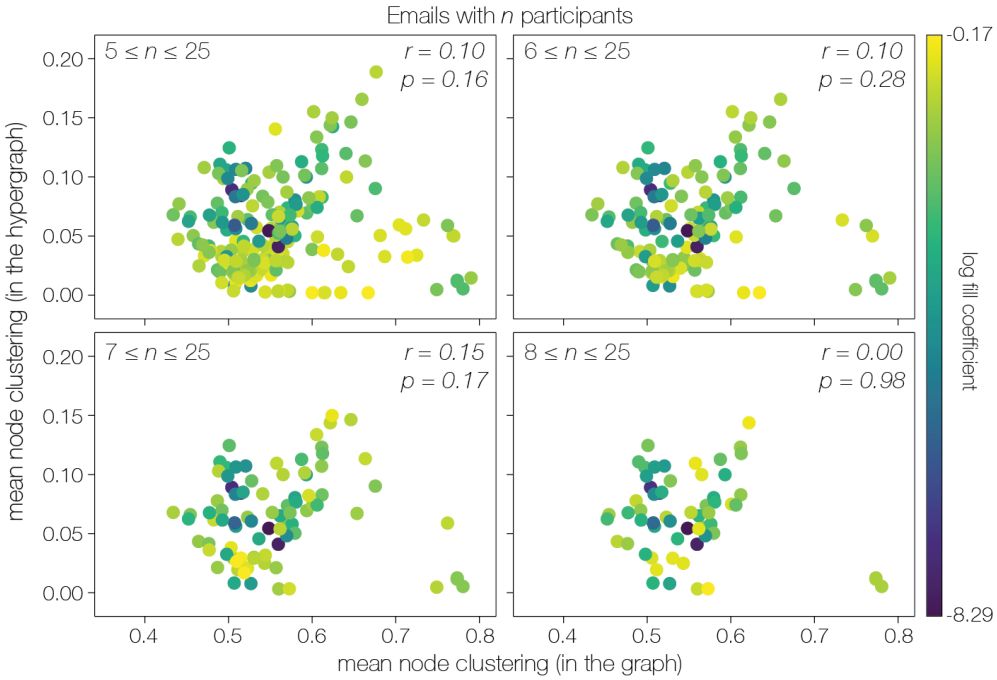

Figure 17: **Graph clustering and hypergraph clustering coefficients are loosely related.** For different ranges of hyperedge cardinality, the average clustering coefficient of nodes within a hyperedge calculated with the projected graph definition is compared to the average clustering coefficient calculated from the hypergraph representation. Scatterplot points are colored by the log(fill coefficient).

this information as color on her scatterplots (Fig. 17). By eye her results show no relationship between a team's cohesiveness with the company (clustering coefficient calculated from the graph or hypergraph representation) and a team's internal cohesiveness (fill coefficient).

Together, these experiments illustrate that even in the case when a) the research question is fixed, b) the researcher has access to the full dataset, and c) there are little-to-no interactions among different types of dependencies, the choice of representation alone may still yield different insights by virtue of the different assumptions made by each (here dyadic *versus* polyadic interactions).

7 Applications

We can naturally encode myriad systems in the real world with at least one of the formalisms discussed in this work. Still, often we focus on analyzing a system from a *particular* perspective and spend less time imagining how alternative analysis pipelines may be more revealing – or indeed more faithful to the system – than the currently used pipeline. Here we consider the alternative perspectives offered by the dependencies, formalisms, and challenges we have discussed in this paper.

First we consider the brain. The brain can be naturally conceived of as a system of individual parts that work together to form large functional units at different scales: neurons work together to communicate with each other forming co-firing patterns called *code words* [54], multiple neuronal populations collaborate in order to plan and evaluate trajectories [145], and entire brain regions work in unison to form functional networks [81]. Scientists have successfully studied the brain by encoding it at any one of these scales using the representations discussed in this work [27]. For example, with graph representations researchers found the brain to exhibit small-worldness [19, 15, 139], modular architecture [196, 81], and hubs [2, 3]. Using the simplicial complex representation, at the larger scale cavities in the structural adult brain were observed [159, 191], and at a smaller scale researchers detected the geometric structure of pyramidal neuron firing patterns [86]. Finally, research employing a hypergraph representation has identified functional hub hyperedges [219], characterized types of hyperedges in developing children [89], and tracked changes in brain organization over both short [18, 56] and long time scales [57].

Looking to the future, one particularly little-understood aspect of the brain is the impact of temporal dependency. How do specific relations affect the existence of any other relations in the future? For example, can a brain transition from any arbitrary state to any other state [49], or is its future activity bound by its past activity [12, 217]? Though the field has used temporal networks to investigate time-varying activity, at the time of writing we did not find the inclusion of temporal dependencies within the representation. We suggest the application of higher-order network representations to deepen our understanding of temporal dependencies in this complex system. Additionally, at all scales the brain is spatially embedded [199], and previous work has shown that the strength of connections between brain regions often depends on the euclidean distance between them [96, 170]. Often the analysis pipeline involves comparing any computed results on the empirical data against a spatially-embedded null model [28], or co-modeling the spatial relations and other relations such as by examining Rentian scaling [16, 97, 177]. While these approaches do help to determine dependence of spatial features on other features, we suggest taking an additional step to directly encoding spatial dependencies in the formal representation of the data. For example, multilayer representations or sheaves encoding position information may be of help here.

Next we consider transportation, which is another well-studied system in network science. Transportation networks come in two different types: systems where the movement is done along fixed routes (such as roads, train tracks, power lines, or airline paths [171, 20, 183, 150, 47]), and systems where the movement is done freely through (outer) space [173]. Among these, analyses specifically of public transportation networks have incorporated multilayer networks [215] or variations thereof including internal node structure [186], and have also evaluated system-specific measures that include spatial organization [216]. Analyses involving temporal representations have included investigating congestion clusters in road networks [168] and how to alleviate them [102]. These analyses usually consider spatial and temporal dependencies by assigning weights to the representation's relations associated to distances or travel times [160, 181]. Importantly, studies are beginning to encode the temporal dependency in the representation itself, as higher order networks have

revealed these temporal dependencies in data from global shipping and web browsing [223].

The subset dependency is studied far less often than other types of dependencies in transportation systems. For example, in a public transport system, if relations are defined among k stations if there exists a route X that stops at all k stations, then certainly any subset of those k stations must also be related by route X. Alternatively a subset dependency may or may not exist within traversed paths. For example, perhaps we observe paths of length k but we do not observe smaller sub-paths. How might the identification or inclusion of subset dependencies within the transportation system improve our ability to prevent system failures or predict future activity? Additionally, we suggest further investigation of polyadic relations in these systems with simplicial complexes or hypergraphs where appropriate, as these representations may elucidate previously hidden system properties.

Finally we consider applications in cellular systems composed of any subset of proteins, genes, regulatory units such as enhancers, epigenetic factors, and more [77, 5]. Most commonly the field studies system fragments such as genetic regulatory networks (GRNs) [60, 218, 70] and protein-protein interaction networks (PPINs) [61, 164]. Unlike the above two examples, this application differs in that only in rare situations can real-world interactions be observed. Consequently, tremendous effort focuses on network reconstruction from data, i.e. inferring the interactions from indirect measurements [144, 213, 4]. Temporal information such as fluctuations in RNA counts [194] and spatial information such as co-localization [127] can be used to reconstruct the network of interactions. Often, for example in the system fragment of proteins and protein complexes, polyadic relations exist and have been encoded using simplicial complexes or hypergraphs [163]. Multilayer representations have also been used for representing multiple biological layers important in disease [92] and for inferring protein function [225].

The difficulties associated with macromolecule-interaction systems create an enticing problem for developing system representations. Since the existence of interactions can rarely be observed directly, perhaps a representation with weights indicating the probability of the existence of each relation might be a useful alternative. In such a case, one might consider studying an ensemble of graphs, simplicial complexes, or hypergraphs instead of only one, and differentiating among them based on the likelihood of each one being a faithful representation of the real system. Additionally, though polyadic relations are known to play an important role in these systems [205], simplicial complex and hypergraph representations are more rarely employed (examples include [73, 187, 111]. Finally, temporal fluctuations are becoming easier to record in these systems [180], which poses the opportunity to directly study temporal dependencies.

8 Discussion and Conclusion

In this work we examined each step of a data analysis pipeline suitable for studying complex systems (Fig. 18). We first discussed system dependencies which can manifest in different flavors including but not limited to temporal, subset, and spatial. We then defined common complex system formalisms and their underlying assumptions, as well as which dependencies they encode. We discussed the mathematical relationships between formalisms, and how information can be lost (or imputed) as we convert data from one formalism to another. Finally we offered analysis examples in order to underscore the importance of dependencies, careful choice of representation, and analysis techniques in studying complex systems.

The main message of our work is that there is no perfect way to analyze a system, and that studying two different systems may require two entirely different pipelines. That is, the modeling decisions made while studying one dataset compiled from a system will not necessarily carry over to another system or, indeed, not even to another dataset extracted from the same system. In contrast, we see many studies apply certain pipelines for seemingly no other reason than because they are common within a certain field. Instead, we recommend that each new system and dataset be individually evaluated and investigated, and each assumption and pipeline decision be made in accordance with the concepts discussed here. More specifically, and following Fig 18, we suggest designing pipelines based on the system and system dependencies, any external dependencies that may be induced by the data type or data collection method, and the limits of the system fragment under study. From there, we suggest choosing a formalism that best fits the data, the research question, and the system itself, even if it requires using a new formalism or extension that is outside

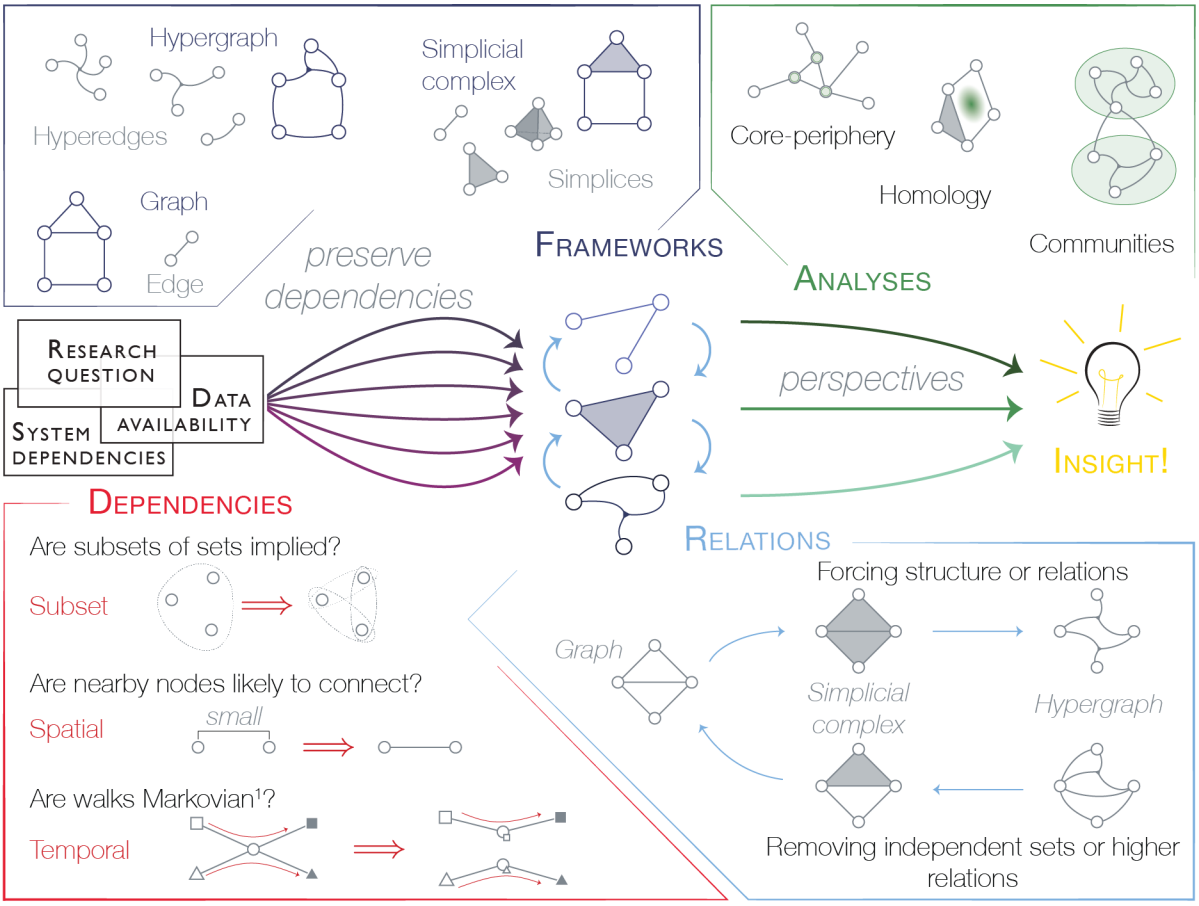

Figure 18: Complex system analysis pipeline discussed in this paper. We suggest beginning with considering how the research question, data availability, and system dependencies may influence downstream analyses. We discussed subset, spatial, and temporal dependencies (red). We suggest choosing a formalism (navy) that preserves and respects dependencies within the system and data. Formalisms themselves are mathematically related (light blue), but switching formalisms after the initial data encoding can result in making inaccurate assumptions or forgetting independent relations. Finally, choosing the appropriate analysis method for the system and representation (green) after all other steps have been performed carefully can offer insight into the system's behavior, structure, or function.

of what is customary. Finally, we recommend choosing carefully the specific methods, measurements, and analyses performed on the chosen representation, and keeping in mind that their results may be biased by the choices made in the previous stages. Different choices at each of these steps may ultimately yield results that are at odds with the results yielded by other choices. Only after respecting the system's dependencies and unique qualities through proper representation and analysis methods will we uncover novel insight into the system under study.

Though here we present only a first attempt to unify the application of complex systems analyses, we hope that the drive for more accurate representations will continue to push the field both forward *and* closer together through multiplying collaborations. We imagine that complex systems researchers in the future may each have a slew of representations along with carefully chosen computations that respect the dependencies one finds within the system. By continuing this discussion, the separate areas of science that use complex systems analyses will together identify what is missing from current formalisms, create more insightful analyses, and generate novel techniques.

9 Acknowledgments

First and foremost, we wish to acknowledge the colleagues, co-authors, mentors, and mentees who have shaped our perspective on this subject. Next, we wish to thank Jason Kim, Linden Parkes, and Shubhankar Patankar for their helpful and constructive comments on an earlier version of this manuscript. Finally, we acknowledge critical financial support that allowed us to devote time to this work. ASB and DSB acknowledge support from the Army Research Office (Falk-W911NF-18-1-0244, Grafton-W911NF-16-1-0474, DCIST- W911NF-17-2-0181), the National Science foundation (PHY-1554488, IIS-1926757), and the Paul G. Allen Family Foundation. LT and TER were supported in part by the National Science Foundation (IIS-1741197) and by the Combat Capabilities Development Command Army Research Laboratory (under Cooperative Agreement Number W911NF-13-2-0045). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation here.

10 Citation diversity statement

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minorities are under-cited relative to the number of such papers in the field [67, 126, 38, 42, 206, 63]. Here we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, and other factors. Gender bias can arise due to explicit and implicit bias against a person's known gender as a woman, or due to explicit or implicit bias against a person carrying a name commonly used by women [123, 152, 137]. To evaluate the former (bias according to known gender), we obtained predicted gender of the first and last author of each reference using pronouns affiliated with them online or pronouns known by personal friendships; by this measure (and excluding self-citations to the first and last authors of our current paper), our references contain 42% man(first)/man(last), 13% man/woman, 10% woman/man, 13% woman/woman, 0% non-binary , and 21% unknown categorization. This method is limited in that pronouns may not be indicative of gender identity, and may not be consistent across time or environment. To evaluate the latter (bias according to a gendered name), we used databases that store the probability of a name being carried by a woman; by this measure (again excluding self-citations), our references contains 60% man/man names, 12% man/woman names, 11% woman/man names, 10% woman/woman names, and 7% unknown categorization [227, 67]. This method is limited in that it cannot account for intersex, non-binary, or transgender people. We look forward to future work that could help us to better understand how to support equitable practices in science.

References

[1] The Apache Software Foundation 2020. Apache airflow. `https://airflow.apache.org/`. Accessed: 2020-05-22.

[2] Sophie Achard, C Delon-Martin, P E Vértes, F Renard, M Schenck, F Schneider, C Heinrich, S Kremer, and Edward T Bullmore. Hubs of brain functional networks are radically reorganized in comatose patients. *Proceedings of the National Academy of Sciences USA*, 109(50):20608–13, 2012.

[3] Sophie Achard, R Salvador, B Whitcher, John Suckling, and Edward Bullmore. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *Journal of Neuroscience*, 26(1):63–72, 2006.

[4] Réka Albert. Network inference, analysis, and modeling in systems biology. *The Plant Cell*, 19(11):3327–3338, 2007.

[5] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

[6] Luis A Nunes Amaral, Antonio Scala, Marc Barthélemy, and H Eugene Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences USA*, 97(21):11149–11152, 2000.

[7] Andrea Apolloni, C Poletto, and Vittoria Colizza. Age-specific contacts and travel patterns in the spatial spread of 2009 H1N1 influenza pandemic. *BMC Infectious Diseases*, 13:176, 2013.

[8] Viplove Arora and Mario Ventresca. Action-based modeling of complex networks. *Scientific Reports*, 7(1):1–10, 2017.

[9] Ronald H Atkin. From cohomology in physics to q-connectivity in social science. *International Journal of Man-Machine Studies*, 4(2):139–167, 1972.

[10] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[11] Matteo Barigozzi, Giorgio Fagiolo, and Giuseppe Mangioni. Identifying the community structure of the international-trade multi-network. *Physica A: Statistical Mechanics and its Applications*, 390(11):2051–2066, 2011.

[12] Anna Barnes, Edward T Bullmore, and John Suckling. Endogenous human brain dynamics recover slowly following cognitive effort. *PLoS One*, 4(8):e6626, 2009.

[13] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011.

[14] Marc Barthélemy. Transitions in spatial networks. *Comptes Rendus Physique*, 19(4):205–232, 2018.

[15] Danielle S Bassett and Edward T Bullmore. Small-world brain networks revisited. *Neuroscientist*, 23(5):499–516, 2017.

[16] Danielle S Bassett, Daniel L Greenfield, Andreas Meyer-Lindenberg, Daniel R Weinberger, Simon W Moore, and Edward T Bullmore. Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Computational Biology*, 6(4):e1000748, 2010.

[17] Danielle S Bassett and Olaf Sporns. Network neuroscience. *Nature Neuroscience*, 20(3):353, 2017.

[18] Danielle S. Bassett, Nicholas F. Wymbs, Mason A. Porter, Peter J. Mucha, and Scott T Grafton. Cross-linked structure of network evolution. *Chaos*, 24(1):013112, 2014.

[19] Danielle Smith Bassett and Edward T Bullmore. Small-world brain networks. *The Neuroscientist*, 12(6):512–523, 2006.

[20] Hannah Bast, Daniel Delling, Andrew Goldberg, Matthias Müller-Hannemann, Thomas Pajor, Peter Sanders, Dorothea Wagner, and Renato F Werneck. Route planning in transportation networks. In *Algorithm Engineering*, pages 19–80. Springer, 2016.

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

[22] Alex Bavelas. Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730, 1950.

[23] Austin R. Benson, Rediet Abebe, Michael T. Schaub, Ali Jadbabaie, and Jon M. Kleinberg. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences USA*, 115(48):E11221–E11230, 2018.

[24] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[25] Austin R Benson, Ravi Kumar, and Andrew Tomkins. Sequences of sets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD*, 2018.

[26] Claude Berge. *Hypergraphs: Combinatorics of Finite Sets*, volume 45. Elsevier, 1984.

[27] Richard F. Betzel and Danielle S. Bassett. Multi-scale brain networks. *NeuroImage*, 160:73–83, 2017.

[28] Richard F. Betzel, John D. Medaglia, Lia Papadopoulos, Graham L Baum, Ruben Gur, Raquel Gur, David Roalf, Theodore D Satterthwaite, and Danielle S Bassett. The modular organization of human anatomical brain networks: Accounting for the cost of wiring. *Network Neuroscience*, 1(1):42–68, 2017.

[29] Ginestra Bianconi. *Multilayer Networks: Structure and Function*. Oxford University Press, 2018.

[30] Jacob Charles Wright Billings, Mirko Hu, Giulia Lerda, Alexey N Medvedev, Francesco Mottes, Adrian Onicas, Andrea Santoro, and Giovanni Petri. Simplex2vec embeddings for community detection in simplicial complexes. *arXiv preprint arXiv:1906.09068*, 2019.

[31] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesús Gómez-Gardenes, Miguel Romance, Irene Sendina-Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

[32] Stephen P Borgatti and Martin G Everett. Models of core/periphery structures. *Social Networks*, 21(4):375–395, 2000.

[33] Christian Borgs and Jennifer Chayes. Graphons: A nonparametric method to model, estimate, and design algorithms for massive networks. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 665–672. ACM, 2017.

[34] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, pages 1–7, 2020.

[35] Charles D Brummitt, Raissa M. D'Souza, and Elizabeth A. Leicht. Suppressing cascades of load in interdependent networks. *Proceddings of the National Academy of Sciences USA*, 109(12):E680–E689, 2012.

[36] Edward T. Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186, 2009.

[37] Carter T Butts. Revisiting the foundations of network analysis. *Science*, 325(5939):414–416, 2009.

[38] Neven Caplar, Sandro Tacchella, and Simon Birrer. Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy*, 1(6):1–5, 2017.

[39] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[40] Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1):1–12, 2008.

[41] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jure Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security*, 10(4):1:1–1:26, 2008.

[42] Paula Chakravartty, Rachel Kuo, Victoria Grubbs, and Charlton McIlwain. #CommunicationSoWhite. *Journal of Communication*, 68(2):254–266, 2018.

[43] Philip Chodrow and Andrew Mellor. Annotated hypergraphs: Models and applications. *Applied Network Science*, 5(1):9, 2020.

[44] Philip S Chodrow, Z Al-Awwad, S Jiang, and Marta C González. Demand and congestion in multiplex transportation networks. *PLoS One*, 11(9):e0161738, 2016.

[45] Phillip Christie and Dirk Stroobandt. The interpretation and application of rent's rule. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 8(6):639–648, 2000.

[46] Aaron Clauset, Daniel B Larremore, and Roberta Sinatra. Data-driven predictions in the science of science. *Science*, 355(6324):477–480, 2017.

[47] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences USA*, 103(7):2015–2020, 2006.

[48] Vittoria Colizza and Alessandro Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *Journal of Theoretical Biology*, 251(3):450–467, 2008.

[49] Eli J. Cornblath, Arian Ashourvan, Jason Z Kim, Richard F. Betzel, Rastko Ciric, Azeez Adebimpe, Graham L Baum, Xiaosong He, Kosha Ruparel, Tyler M Moore, Ruben C Gur, Raquel E Gur, Russel T Shinohara, David R Roalf, Theodore D Satterthwaite, and Danielle S Bassett. Temporal sequences of brain activity at rest are constrained by white matter structure and modulated by cognitive demands. *Communications Biology*, 3(1):261, 2020.

[50] L da F Costa, Francisco A Rodrigues, Gonzalo Travieso, and Paulino Ribeiro Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[51] O T Courtney and G Bianconi. Dense power-law networks and simplicial complexes. *Physical Review E*, 97(5-1):052303, 2018.

[52] Justin M. Curry. *Sheaves, Cosheaves and Their Applications*. PhD thesis, University of Pennsylvania, 2014.

[53] Carina Curto. What can topology tell us about the neural code? *Bulletin of the American Mathematical Society*, 54(1):63–78, 2017.

[54] Carina Curto, Elizabeth Gross, Jack Jeffries, Katherine Morrison, Mohamed Omar, Zvi Rosen, Anne Shiu, and Nora Youngs. What makes a neural code convex? *SIAM Journal on Applied Algebra and Geometry*, 1(1):222–238, 2017.

[55] David B Damiano and Melissa R McGuirl. A topological analysis of targeted in-111 uptake in spect images of murine tumors. *Journal of Mathematical Biology*, 76(6):1559–1587, 2018.

[56] Elizabeth N Davison, Kimberly J Schlesinger, Danielle S Bassett, Mary-Ellen Lynall, Michael B Miller, Scott T Grafton, and Jean M Carlson. Brain network adaptability across task states. *PLoS Computational Biology*, 11(1):e1004029., 2015.

[57] Elizabeth N Davison, B O Turner, Kimberly J Schlesinger, Michael B Miller, Scott T Grafton, Danielle S Bassett, and Jean M Carlson. Individual differences in dynamic functional brain connectivity across the human lifespan. *PLoS Computational Biology*, 12(11):e1005178, 2016.

[58] Manlio De Domenico, Clara Granell, Mason A Porter, and Alex Arenas. The physics of spreading processes in multilayer networks. *Nature Physics*, 12(10):901, 2016.

[59] Manlio De Domenico, Andrea Lancichinetti, Alex Arenas, and Martin Rosvall. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Physical Review X*, 5(1):011027, 2015.

[60] Hidde De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.

[61] Javier De Las Rivas and Celia Fontanillo. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Computational Biology*, 6(6), 2010.

[62] Andrea De Montis, Marc Barthélemy, Alessandro Chessa, and Alessandro Vespignani. The structure of interurban traffic: a weighted network analysis. *Environment and Planning B: Planning and Design*, 34(5):905–924, 2007.

[63] Michelle L Dion, Jane Lawrence Sumner, and Sara McLaughlin Mitchell. Gendered citation patterns across political science and social science methodology fields. *Political Analysis*, 26(3):312–327, 2018.

[64] Clifford H Dowker. Homology groups of relations. *Annals of Mathematics*, pages 84–95, 1952.

[65] Raissa M. D'Souza, Jesus Gómez-Gardeñes, Jan Nagler, and Alex Arenas. Explosive phenomena in complex networks. *Advances in Physics*, 68(3):123–223, 2019.

[66] Olga Dunaeva, Herbert Edelsbrunner, Anton Lukyanov, Michael Machin, Daria Malkova, Roman Kuvaev, and Sergey Kashin. The classification of endoscopy images with persistent homology. *Pattern Recognition Letters*, 83:13–22, 2016.

[67] Jordan D Dworkin, Kristin A Linn, Erin G Teich, Perry Zurn, Russell T Shinohara, and Danielle S Bassett. The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*, 23, 2020.

[68] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Proceedings. 41st Annual Symposium on Foundations of Computer Science*, pages 454–463. IEEE, 2000.

[69] Daniel Edler, Ludvig Bohlin, et al. Mapping higher-order network flows in memory and multilayer networks with infomap. *Algorithms*, 10(4):112, 2017.

[70] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in Cell and Developmental Biology*, 2:38, 2014.

[71] Ernesto Estrada, Gissell Estrada-Rodriguez, and Heiko Gimperlein. Metaplex networks: influence of the exo-endo structure of complex systems on diffusion. *SIAM Review*, 62(3), 2020.

[72] Ernesto Estrada and Juan A Rodríguez-Velázquez. Subgraph centrality and clustering in complex hyper-networks. *Physica A: Statistical Mechanics and its Applications*, 364:581–594, 2006.

[73] Ernesto Estrada and Grant J Ross. Centralities in simplicial complexes. Applications to protein interaction networks. *Journal of Theoretical Biology*, 438:46–60, 2018.

[74] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140, 1741.

[75] Shimon Even. *Graph Algorithms*. Cambridge University Press, 2011.

[76] Paul Expert, Tim S Evans, Vincent D Blondel, and Renaud Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences USA*, 108(19):7663–7668, 2011.

[77] Valeria Fionda. *Networks in Biology*. Elsevier, 2019.

[78] Linton Freeman. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical press, 2004.

[79] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[80] Suzanne Renick Gallagher and Debra S Goldberg. Clustering coefficients in protein interaction hypernetworks. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, pages 552–560, 2013.

[81] C L Gallen and M D'Esposito. Brain modularity: A biomarker of intervention-related plasticity. *Trends in Cognitive Science*, 23(4):293–304, 2019.

[82] Giorgio Gallo, Giustino Longo, Stefano Pallottino, and Sang Nguyen. Directed hypergraphs and applications. *Discrete Applied Mathematics*, 42(2-3):177–201, 1993.

[83] Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[84] Robert Ghrist and Yasuaki Hiraoka. Applications of sheaf cohomology and exact sequences to network coding. In *Proceedings of the 2011 Nonlinear Theory and its Applications*. NOLTA, 2011.

[85] Chad Giusti, Robert Ghrist, and Danielle S Bassett. Two's company, three (or more) is a simplex. *Journal of computational neuroscience*, 41(1):1–14, 2016.

[86] Chad Giusti, Eva Pastalkova, Carina Curto, and Vladimir Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences USA*, 112(44):13455–13460, 2015.

[87] Stefania Gnesi, Ugo Montanari, and Alberto Martelli. Dynamic programming as graph searching: An algebraic approach. *Journal of the ACM (JACM)*, 28(4):737–751, 1981.

[88] Deanna J Greene, Christina N Lessov-Schlaggar, and Bradley L Schlaggar. Development of the brain's functional network architecture. In *Neurobiology of Language*, pages 399–406. Elsevier, 2016.

[89] Shi Gu, Muzhi Yang, John D Medaglia, Ruben C Gur, Raquel E Gur, Theodore D Satterthwaite, and Danielle S Bassett. Functional hypergraph uncovers novel covariant structures over neurodevelopment. *Human Brain Mapping*, 38(8):3823–3835, 2017.

[90] Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101, 2004.

[91] Mahnaz Habibi and Pegah Khosravi. Disruption of the protein complexes from weighted complex networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018.

[92] Arda Halu, Manlio De Domenico, Alex Arenas, and Amitabh Sharma. The multiplex network of human diseases. *NPJ Systems Biology and Applications*, 5(1):1–12, 2019.

[93] Ilkka Hanski. *Metapopulation Ecology*. Oxford University Press, 1999.

[94] Frank Harary and Robert Z Norman. Some properties of line digraphs. *Rendiconti del Circolo Matematico di Palermo*, 9(2):161–168, 1960.

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

[96] Szabolcs Horvát, Răzvan Gămănuț, Mária Ercsey-Ravasz, Loïc Magrou, Bianca Gămănuț, David C Van Essen, Andreas Burkhalter, Kenneth Knoblauch, Zoltán Toroczkai, and Henry Kennedy. Spatial embedding and wiring cost constrain the functional layout of the cortical network of rodents and primates. *PLoS Biology*, 14(7):e1002512, 2016.

[97] Javier J How and Saket Navlakha. Evidence of rentian scaling of functional modules in diverse biological networks. *Neural Computation*, 30(8):2210–2244, 2018.

[98] Sen Hu, Hualei Yang, Boliang Cai, and Chunxia Yang. Research on spatial economic structure for different economic sectors from a perspective of a complex network. *Physica A: Statistical Mechanics and its Applications*, 392(17):3682–3697, 2013.

[99] Iacopo Iacopini, Giovanni Petri, Alain Barrat, and Vito Latora. Simplicial models of social contagion. *Nature Communications*, 10(1):1–9, 2019.

[100] Paul Samuel P Ignacio and Isabel K Darcy. Tracing patterns and shapes in remittance and migration networks via persistent homology. *EPJ Data Science*, 8(1):1, 2019.

[101] Bukyoung Jhun, Minjae Jo, and B Kahng. Simplicial sis model in scale-free uniform hypergraph. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):123207, 2019.

[102] Yuxuan Ji and Nikolas Geroliminis. Spatial and temporal analysis of congestion in urban transportation networks. In *Transportation Research Board Annual Meeting*. Washington DC, 2011.

[103] Christopher W Johnson. What are emergent properties and how do they affect the engineering of complex systems? *Reliability Engineering and System Safety*, 91(12):1475–1481, 2006.

[104] Matthew Kahle. Topology of random clique complexes. *Discrete Mathematics*, 309(6):1658–1671, 2009.

[105] Matthew Kahle. Sharp vanishing thresholds for cohomology of random flag complexes. *Annals of Mathematics*, pages 1085–1107, 2014.

[106] Ankit N Khambhati, Ann E Sizemore, Richard F Betzel, and Danielle S Bassett. Modeling and interpreting mesoscale network dynamics. *NeuroImage*, 180:337–349, 2018.

[107] Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Community detection in hypergraphs, spiked tensor models, and sum-of-squares. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 124–128. IEEE, 2017.

[108] Mikko Kivelä, Alex Arenas, Marc Barthélemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

[109] Sophia Kivelson and Steven A Kivelson. Defining emergence in physics. *Nature Partner Journals Quantum Materials*, 1:16024, 2016.

[110] Steffen Klamt and Ernst Dieter Gilles. Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2):226–234, 2004.

[111] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS Computational Biology*, 5(5), 2009.

[112] Michael Kotliar, Andrey V Kartashov, and Artem Barski. Cwl-airflow: a lightweight pipeline manager supporting common workflow language. *GigaScience*, 8(7):giz084, 2019.

[113] Larkshmi Krishnamurthy, J Nadeau, Gultekin Ozsoyoglu, M Ozsoyoglu, Greg Schaeffer, Murat Tasan, and Wanhong Xu. Pathways database system: an integrated system for biological pathways. *Bioinformatics*, 19(8):930–937, 2003.

[114] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences USA*, 110(52):20935–20940, 2013.

[115] James Ladyman and Karoline Wiesner. *What is a complex system?* Yale University Press, 2020.

[116] Eric C Lai. Notch signaling: control of cell communication and cell fate. *Development*, 131(5):965–973, 2004.

[117] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes. From networks to optimal higher-order models of complex systems. *Nature Physics*, page 1, 2019.

[118] Renaud Lambiotte, Vsevolod Salnikov, and Martin Rosvall. Effect of memory on the dynamics of random walks on networks. *Journal of Complex Networks*, 3(2):177–188, 2014.

[119] Giorgio Levi and Franco Sirovich. Generalized and/or graphs. *Artificial Intelligence*, 7(3):243–259, 1976.

[120] Richard Levins. Some demographic and genetic consequences of environmental heterogeneity for biological control. *American Entomologist*, 15(3):237–240, 1969.

[121] Jiaqi Liang, L Li, and Daniel Zeng. Evolutionary dynamics of cryptocurrency transaction networks: An empirical study. *PLoS One*, 13(8):e0202202, 2018.

[122] Antonio Lima, R Stanojevic, D Papagiannaki, P Rodriguez, and Marta C González. Understanding individual routing behaviour. *Journal of the Royal Society Interface*, 13(116):20160021, 2016.

[123] Lillian MacNell, Adam Driscoll, and Andrea N Hunt. What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4):291–303, 2015.

[124] Parul Maheshwari, H Du, J Sheen, S M Assmann, and Reka Albert. Model-driven discovery of calcium-related protein-phosphatase inhibition in plant guard cell signaling. *PLoS Computational Biology*, 15(10):e1007429, 2019.

[125] Slobodan Maletić, Milan Rajković, and Danijela Vasiljević. Simplicial complexes of networks and their statistical properties. In *International Conference on Computational Science*, pages 568–575. Springer, 2008.

[126] Daniel Maliniak, Ryan Powers, and Barbara F Walter. The gender citation gap in international relations. *International Organization*, 67(4):889–922, 2013.

[127] Faraz K Mardakheh, Heba Z Sailem, Sandra Kümper, Christopher J Tape, Ryan R McCully, Angela Paul, Sara Anjomani-Virmouni, Claus Jørgensen, George Poulogiannis, Christopher J Marshall, et al. Proteomics profiling of interactome dynamics by colocalisation analysis (cola). *Molecular BioSystems*, 13(1):92–105, 2017.

[128] Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559, 2005.

[129] Giulia Menichetti, L Dall'Asta, and Ginestra Bianconi. Control of multilayer networks. *Sci Rep*, 6:20706, 2016.

[130] Giulia Menichetti, Daniel Remondini, Pietro Panzarasa, Raúl J Mondragón, and Ginestra Bianconi. Weighted multiplex networks. *PloS One*, 9(6):e97857, 2014.

[131] Nikola Milosavljević, Dmitriy Morozov, and Primoz Skraba. Zigzag persistent homology in matrix multiplication time. In *Proceedings of the Twenty-Seventh Annual Symposium on Computational Geometry*, pages 216–225. ACM, 2011.

[132] Melanie Mitchell. Complex systems: Network thinking. *Artificial Intelligence*, 170(18):1194–1212, 2006.

[133] Melanie Mitchell. *Complexity: A Guided Tour*. Oxford University Press, 2009.

[134] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.

[135] José M Montoya, Stuart L Pimm, and Ricard V Solé. Ecological networks and their fragility. *Nature*, 442(7100):259, 2006.

[136] Katherine Morrison and Carina Curto. Predicting neural network dynamics via graphical analysis. In *Algebraic and Combinatorial Computational Biology*, pages 241–277. Elsevier, 2019.

[137] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences USA*, 109(41):16474–16479, 2012.

[138] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[139] Sarah F. Muldoon, Eric W. Bridgeford, and Danielle S. Bassett. Small-world propensity and weighted brain networks. *Scientific Reports*, 6:22057, 2016.

[140] Andrew C Murphy, Shi Gu, Ankit N Khambhati, Nicholas F Wymbs, Scott T Grafton, Theodore D Satterthwaite, and Danielle S Bassett. Explicitly linking regional activation and function connectivity: community structure of weighted networks with continuous annotation. *arXiv preprint arXiv:1611.07962*, 2016.

[141] Nicolas Neubauer and Klaus Obermayer. Towards community detection in k-partite k-uniform hypergraphs. In *Proceedings of the NIPS 2009 Workshop on Analyzing Networks and Learning with Graphs*, pages 1–9, 2009.

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

[143] Vincenzo Nicosia, John Tang, Cecilia Mascolo, Mirco Musolesi, Giovanni Russo, and Vito Latora. Graph metrics for temporal networks. In *Temporal Networks*, pages 15–40. Springer, 2013.

[144] Chris J Oates and Sach Mukherjee. Network inference and biological dynamics. *The Annals of Applied Statistics*, 6(3):1209, 2012.

[145] H Freyja Ólafsdóttir, Daniel Bush, and Caswell Barry. The role of hippocampal replay in memory and planning. *Current Biology*, 28(1):R37–R50, 2018.

[146] Jukka-Pekka Onnela, Daniel J Fenn, Stephen Reid, Mason A Porter, Peter J Mucha, Mark D Fricker, and Nick S Jones. Taxonomies of networks from community structure. *Physical Review E*, 86(3):036104, 2012.

[147] Edward Ott, J H Platig, T M Antonsen, and Michelle Girvan. Echo phenomena in large systems of coupled oscillators. *Chaos*, 18(3):037115, 2008.

[148] Nina Otter, Mason A Porter, U Tillmann, P Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):17, 2017.

[149] Can Özturan. On finding hypercycles in chemical reaction networks. *Applied Mathematics Letters*, 21(9):881–884, 2008.

[150] Giuliano Andrea Pagani and Marco Aiello. The power grid as a complex network: a survey. *Physica A: Statistical Mechanics and its Applications*, 392(11):2688–2700, 2013.

[151] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814, 2005.

[152] Michele A Paludi and Lisa A Strayer. What's in an author's name? differential evaluations of performance as a function of author's name. *Sex Roles*, 12(3-4):353–361, 1985.

[153] Joshua Pan, Robin M Meyers, Brittany C Michel, Nazar Mashtalir, Ann E Sizemore, Jonathan N Wells, Seth H Cassel, Francisca Vazquez, Barbara A Weir, William C Hahn, et al. Interrogation of mammalian protein complex structure, function, and membership using genome-scale fitness screens. *Cell Systems*, 6(5):555–568, 2018.

[154] Lia Papadopoulos, Pablo Blinder, Henrik Ronellenfitsch, Florian Klimm, Eleni Katifori, David Kleinfeld, and Danielle S Bassett. Comparing two classes of biological distribution systems using network analysis. *PLoS Computational Biology*, 14(9):e1006428, 2018.

[155] Lia Papadopoulos, Jason Z Kim, Jürgen Kurths, and Danielle S Bassett. Development of structural correlations and synchronization from adaptive rewiring in networks of kuramoto oscillators. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(7):073115, 2017.

[156] Lia Papadopoulos, Mason A Porter, Karen E Daniels, and Danielle S Bassett. Network analysis of particles and grains. *Journal of Complex Networks*, 6(4):485–565, 2018.

[157] Jorge Peña and Yannick Rochat. Bipartite graphs as models of population structures in evolutionary multiplayer games. *PloS One*, 7(9), 2012.

[158] Vincenzo Perri and Ingo Scholtes. Higher-order visualization of causal structures in dynamics graphs. *arXiv preprint arXiv:1908.05976*, 2019.

[159] Giovanni Petri, Paul Expert, Federico Turkheimer, Robin Carhart-Harris, David Nutt, Peter J Hellyer, and Francesco Vaccarino. Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface*, 11(101):20140873, 2014.

[160] Sergio Porta, Paolo Crucitti, and Vito Latora. The network analysis of urban streets: a primal approach. *Environment and Planning B: Planning and Design*, 33(5):705–725, 2006.

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

[162] Emilie Purvine, Sinan Aksoy, Cliff Joslyn, Kathleen Nowak, Brenda Praggastis, and Michael Robinson. A topological approach to representational data models. In *International Conference on Human Interface and the Management of Information*, pages 90–109. Springer, 2018.

[163] Emad Ramadan, Arijit Tarafdar, and Alex Pothen. A hypergraph model for the yeast protein complex network. In *IEEE 18th International Parallel and Distributed Processing Symposium*, page 189, 2004.

[164] Karthik Raman. Construction and analysis of protein–protein interaction networks. *Automated Experimentation*, 2(1):2, 2010.

[165] Damien Ramel, Xiaobo Wang, Carl Laflamme, Denise J Montell, and Gregory Emery. Rab11 regulates cell–cell communication during collective cell movements. *Nature Cell Biology*, 15(3):317–324, 2013.

[166] Matthew G Rees, Brinton Seashore-Ludlow, and Paul A Clemons. Computational analyses connect small-molecule sensitivity to cellular features using large panels of cancer cell lines. In *Systems Chemical Biology*, pages 233–254. Springer, 2019.

[167] Michael W Reimann, Max Nolte, Martina Scolamiero, Katharine Turner, Rodrigo Perin, Giuseppe Chindemi, Paweł Dłotko, Ran Levi, Kathryn Hess, and Henry Markram. Cliques of neurons bound into cavities provide a missing link between structure and function. *Frontiers in Computational Neuroscience*, 11:48, 2017.

[168] Felix Rempe, Gerhard Huber, and Klaus Bogenberger. Spatio-temporal congestion patterns in urban traffic networks. *Transportation Research Procedia*, 15:513–524, 2016.

[169] Soufiane Rital, Hocine Cherifi, and Serge Miguet. Weighted adaptive neighborhood hypergraph partitioning for image segmentation. In *International Conference on Pattern Recognition and Image Analysis*, pages 522–531. Springer, 2005.

[170] Marta Rivera-Alba, Shiv N Vitaladevuni, Yuriy Mishchenko, Zhiyuan Lu, Shin-ya Takemura, Lou Scheffer, Ian A Meinertzhagen, Dmitri B Chklovskii, and Gonzalo G de Polavieja. Wiring economy and volume exclusion determine neuronal placement in the drosophila brain. *Current Biology*, 21(23):2000–2005, 2011.

[171] Jean-Paul Rodrigue. *The geography of transport systems*. Taylor & Francis, 2016.

[172] M Puck Rombach, Mason A Porter, James H Fowler, and Peter J Mucha. Core-periphery structure in networks. *SIAM Journal on Applied Mathematics*, 74(1):167–190, 2014.

[173] Shane Ross and Martin Lo. The lunar l1 gateway-portal to the stars and beyond. In *AIAA Space 2001 Conference and Exposition*, page 4768, 2001.

[174] Martin Rosvall, Alcides V Esquivel, Andrea Lancichinetti, Jevin D West, and Renaud Lambiotte. Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*, 5:4630, 2014.

[175] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010.

[176] Assieh Saadatpour and Reka Albert. Discrete dynamic modeling of signal transduction networks. *Methods Mol Biol*, 880:255–272, 2012.

[177] A J Sadovsky and J N MacLean. Mouse visual neocortex supports multiple stereotyped patterns of microcircuit activity. *Journal of Neuroscience*, 34(23):7769–7777, 2014.

[178] Faryad Darabi Sahneh and Caterina Scoglio. Competitive epidemic spreading over arbitrary multilayer networks. *Physical Review E*, 89(6):062817, 2014.

[179] Mostafa Salehi, Rajesh Sharma, Moreno Marzolla, Matteo Magnani, Payam Siyari, and Danilo Montesi. Spreading processes in multilayer networks. *IEEE Transactions on Network Science and Engineering*, 2(2):65–83, 2015.

[180] Anca F Savulescu, Robyn Brackin, Emmanuel Bouilhol, Benjamin Dartigues, Jonathan H Warrell, Mafalda R Pimentel, Stephane Dallongeville, Jan Schmoranzer, Jean-Christophe Olivo-Marin, Edgar R Gomes, et al. Dypfish: Dynamic patterned fish to interrogate rna and protein spatial and temporal subcellular distribution. *bioRxiv*, page 536383, 2019.

[181] Jan Scheurer, Carey Curtis, and Sergio Porta. *Spatial network analysis of public transport systems: Developing a strategic planning tool to assess the congruence of movement and urban structure in Australian cities*. GAMUT, Australasian Centre for the Governance and Management of Urban Transport, 2008.

[182] Ingo Scholtes. When is a network a network? Multi-order graphical model selection in pathways and temporal networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD*, 2017.

[183] Katherine A Seaton and Lisa M Hackett. Stations, trains and small-world networks. *Physica A: Statistical Mechanics and its Applications*, 339(3-4):635–644, 2004.

[184] Stephen B Seidman. Structures induced by collections of subsets: A hypergraph approach. *Mathematical Social Sciences*, 1(4):381–396, 1981.

[185] Daniel Hernández Serrano, Juan Hernández-Serrano, and Darío Sánchez Gómez. Simplicial degree in complex networks. applications of topological data analysis to network science. *Chaos, Solitons & Fractals*, 137:109839, 2020.

[186] Tanuja Shanmukhappa, Ivan WH Ho, K Tse Chi, Xingtang Wu, and Hairong Dong. Multi-layer public transport network analysis. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2018.

[187] Daniel Shnier, Mircea A Voineagu, and Irina Voineagu. Persistent homology analysis of brain transcriptome data in autism. *Journal of the Royal Society Interface*, 16(158):20190531, 2019.

[188] R Sinatra, D Condorelli, and V Latora. Networks of motifs from sequences of symbols. *Physical Review Letters*, 105(17):178702, 2010.

[189] Gurjeet Singh, Facundo Memoli, Tigran Ishkhanov, Guillermo Sapiro, Gunnar Carlsson, and Dario L Ringach. Topological analysis of population activity in visual cortex. *Journal of Vision*, 8(8):11–11, 2008.

[190] Ann E Sizemore and Danielle S Bassett. Dynamic graph metrics: Tutorial, toolbox, and tale. *NeuroImage*, 180:417–427, 2018.

[191] Ann E Sizemore, Chad Giusti, Ari Kahn, Jean M Vettel, Richard F Betzel, and Danielle S Bassett. Cliques and cavities in the human connectome. *Journal of Computational Neuroscience*, 44(1):115–145, 2018.

[192] Ann E Sizemore, Elisabeth A Karuza, Chad Giusti, and Danielle S Bassett. Knowledge gaps in the early growth of semantic feature networks. *Nature Human Behaviour*, 2(9):682, 2018.

[193] Luis Solá, Miguel Romance, Regino Criado, Julio Flores, Alejandro García del Amo, and Stefano Boccaletti. Eigenvector centrality of nodes in multiplex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(3):033131, 2013.

[194] Daniel Spies and Constance Ciaudo. Dynamics in transcriptomics: advancements in rna-seq time course and downstream analysis. *Computational and Structural Biotechnology Journal*, 13:469–477, 2015.

[195] David I Spivak. Higher-dimensional models of networks. *arXiv preprint arXiv:0909.4314*, 2009.

[196] Olaf Sporns and Richard F Betzel. Modular brain networks. *Annual Review of Psychology*, 67:613–640, 2016.

[197] Shane Squires, K Sytwu, D Alcala, T M Antonsen, Edward Ott, and Michelle Girvan. Weakly explosive percolation in directed networks. *Physical Review E*, 87(5):052127, 2013.

[198] Massimo Stella, Nicole M Beckage, Markus Brede, and Manlio De Domenico. Multiplex model of mental lexicon reveals explosive learning in humans. *Scientific Reports*, 8(1):2259, 2018.

[199] Jennifer Stiso and Danielle S Bassett. Spatial embedding imposes constraints on neuronal network architectures. *Trends in Cognitive Sciences*, 2018.

[200] Jennifer Stiso, Ankit N Khambhati, Tommaso Menara, Ari E Kahn, Joel M Stein, Sandihitsu R Das, Richard Gorniak, Joseph Tracy, Brian Litt, Kathryn A Davis, et al. White matter network architecture guides direct electrical stimulation through optimal state transitions. *Cell Reports*, 28(10):2554–2566, 2019.

[201] Bernadette Stolz. Computational topology in neuroscience. *Master's thesis (University of Oxford, 2014). Google Scholar*, 2014.

[202] Bernadette J Stolz, Tegan Emerson, Satu Nahkuri, Mason A Porter, and Heather A Harrington. Topological data analysis of task-based fmri data from experiments on schizophrenia. *arXiv preprint arXiv:1809.08504*, 2018.

[203] Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella Pérez, Matteo Caorsi, Anibal Medina-Mardones, Alberto Dassatti, and Kathryn Hess. giotto-tda: A topological data analysis toolkit for machine learning and data exploration. *arXiv preprint arXiv:2004.02551*, 2020.

[204] Caz M Taylor and Richard J Hall. Metapopulation models for seasonally migratory animals. *Biology Letters*, 8(3):477–480, 2011.

[205] Matthew B Taylor and Ian M Ehrenreich. Higher-order genetic interactions and their contribution to complex traits. *Trends in Genetics*, 31(1):34–40, 2015.

[206] Yannik Thiem, Kris F Sealey, Amy E Ferrer, Adriel M Trott, and Rebecca Kennison. Just ideas? the status and future of publication ethics in philosophy: A white paper. Technical report, Technical report, 2018.

[207] Ze Tian, TaeHyun Hwang, and Rui Kuang. A hypergraph-based learning algorithm for classifying gene expression and arraycgh data with prior knowledge. *Bioinformatics*, 25(21):2831–2838, 2009.

[208] Michele Tizzoni, P Bajardi, A Decuyper, G Kon Kam King, C M Schneider, V Blondel, Z Smoreda, Marta C González, and Vittoria Colizza. On the use of human mobility proxies for modeling epidemics. *PLoS Computational Biology*, 10(7):e1003716, 2014.

[209] Eugenio Valdano, Luca Ferreri, Chiara Poletto, and Vittoria Colizza. Analytical computation of the epidemic threshold on temporal networks. *Physical Review X*, 5(2):021005, 2015.

[210] Paola Valdivia, Paolo Buono, and Jean-Daniel Fekete. Hypenet: Visualizing dynamic hypergraphs. In *EuroVis 2017-19th EG/VGC Conference on Visualization*, pages 1–3, 2017.

[211] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.

[212] Nele Vandersickel, Enid Van Nieuwenhuyse, Nicolas Van Cleemput, Jan Goedgebeur, Milad El Haddad, Jan De Neve, Anthony Demolder, Teresa Strisciuglio, Mattias Duytschaever, and Alexander Panfilov. Directed networks as a novel way to describe and analyze cardiac excitation: Directed graph mapping. *Frontiers in physiology*, 10:1138, 2019.

[213] Giuseppe Vinci, Gautam Dasarathy, and Genevera I Allen. Graph quilting: graphical model selection from partially observed covariances. *arXiv preprint arXiv:1912.05573*, 2019.

[214] Vitaly Ivanovich Voloshin. *Introduction to Graph and Hypergraph Theory*. Nova Science Publishers Hauppauge, 2009.

[215] Christian Von Ferber, Taras Holovatch, Yu Holovatch, and V Palchykov. Public transport networks: empirical analysis and modeling. *The European Physical Journal B*, 68(2):261–275, 2009.

[216] Christian von Ferber, Taras Holovatch, Yu Holovatch, and Vasyl Palchykov. Network harness: Metropolis public transport. *Physica A: Statistical Mechanics and its Applications*, 380:585–591, 2007.

[217] F von Wegner, E Tagliazucchi, and H Laufs. Information-theoretical analysis of resting state EEG microstate sequences - non-Markovianity, non-stationarity and periodicities. *NeuroImage*, 158:99–111, 2017.

[218] Albertha JM Walhout. Gene-centered regulatory network mapping. In *Methods in Cell Biology*, volume 106, pages 271–288. Elsevier, 2011.

[219] Zhijiang Wang, Jiming Liu, Ning Zhong, Yulin Qin, Haiyan Zhou, Jian Yang, and Kuncheng Li. A naive hypergraph model of brain networks. In *International Conference on Brain Informatics*, pages 119–129. Springer, 2012.

[220] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440, 1998.

[221] Joe Weber and Mei-Po Kwan. Bringing time back in: A study on the influence of travel time variations and facility opening hours on individual accessibility. *The Professional Geographer*, 54(2):226–240, 2002.

[222] Robin J Wilson. History of graph theory. In *Handbook of Graph Theory*, pages 31–51. Chapman and Hall/CRC, 2013.

[223] Jian Xu, Thanuka L Wickramarathne, and Nitesh V Chawla. Representing higher-order dependencies in networks. *Science Advances*, 2(5):e1600028, 2016.

[224] Jaejun Yoo, Eun Young Kim, Yong Min Ahn, and Jong Chul Ye. Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages. *Journal of Neuroscience Methods*, 267:1–13, 2016.

[225] Bihai Zhao, Sai Hu, Xueyong Li, Fan Zhang, Qinglong Tian, and Wenyin Ni. An efficient method for protein function annotation based on multilayer protein networks. *Human genomics*, 10(1):33, 2016.

[226] Yaofeng D Zhong, V Srivastava, and Naomi E. Leonard. On the linear threshold model for diffusion of innovations in multiplex social networks. *IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2593–2598, 2017.

[227] Dale Zhou, Eli J. Cornblath, Jennifer Stiso, Erin G. Teich, Jordan D. Dworkin, Ann S. Blevins, and Danielle S. Bassett. Gender diversity statement and code notebook v1.0, February 2020.

[228] Wanding Zhou and Luay Nakhleh. Properties of metabolic graphs: biological organization or representation artifacts? *BMC Bioinformatics*, 12(1):132, 2011.

[229] Mengxiao Zhu and Mo Zhang. Network analysis of conversation data for engineering professional skills assessment. *ETS Research Report Series*, 2017(1):1–13, 2017.

[230] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.