Network Structure Inference, A Survey: Motivations, Methods, and Applications

Ivan Brugere, University of Illinois at Chicago
Brian Gallagher, Lawrence Livermore National Laboratory
Tanya Y. Berger-Wolf, University of Illinois at Chicago

Networks are used to represent relationships between entities in many complex systems, spanning from online social networks to biological cell development and brain activity. These networks model relationships which present various challenges. In many cases, relationships between entities are unambiguously known: are two users friends in a social network? Do two researchers collaborate on a published paper? Do two road segments in a transportation system intersect? These are unambiguous and directly observable in the system in question. In most cases, relationship between nodes are not directly observable and must be inferred: does one gene regulate the expression of another? Do two animals who physically co-locate have a social bond? Who infected whom in a disease outbreak?

Existing approaches use specialized knowledge in different home domains to infer and measure the goodness of inferred network for a specific task. However, current research lacks a rigorous validation framework which employs standard statistical validation. In this survey, we examine how network representations are learned from non-network data, the variety of questions and tasks on these data over several domains, and validation strategies for measuring the inferred network's capability of answering questions on the original system of interest.

1. INTRODUCTION

> "Networks are too easy to create, and too difficult to validate" – Ancient Proverb

Networks are used to represent relationships between entities in many complex systems, spanning from online social networks to biological cell development and brain activity. These networks model relationships which present various challenges. In many cases, relationships between entities are unambiguously known: are two users friends in a social network? Do two researchers collaborate on a published paper? Do two road segments in a transportation system intersect? These are unambiguous and directly observable in the system in question [Kramer et al. 2009]. In *most* cases, relationship between nodes are not directly observable and must be inferred: does one gene regulate the expression of another? Do two animals who physically co-locate have a social bond? Who infected whom in a disease outbreak?

Networks are mathematical representations (i.e. models) used to answer these types of questions about data collected on individual entities. There are a broad range of the questions asked, and a variety of ways in which networks are used to answer these questions. However, how do we know if a particular network representation of the data is the most useful in answering a given question? What is the "right" network representation, and how do we compare the utility of many possible representations for our particular question? Finally, how can we measure whether a network is the appropriate model to answer a question of interest on the original system or data?

Existing approaches use specialized knowledge in different home domains to infer and measure the goodness of inferred network for a specific task. Current research lacks a rigorous validation framework which employs standard statistical validation: significance/uniqueness ("How unique is a well-performing network in the space of possible solutions?"), sensitivity ("how does the performance of the network change to changes in the underlying data measurement or model parameters?"), and robustness ("how accurate is this network over different methods or tasks?").

In this survey, we examine how network representations are learned from non-network data, the variety of questions and tasks on these data over several domains, and validation strategies for measuring the inferred network's capability of answering questions on the original system of interest.

1.1. Motivation: Networks Model Complex Relationships

Networks are a natural choice of data representation across many domains. Networks naturally represent *higher-order structure* which emerge from dyadic relationships, and serve as units of further analysis. These structures include neighborhoods, ego-nets, communities/modules, and connected components. For example, in the computational biology domain, clusters and motifs often represent shared biological roles of proteins. The collective connectivity over these proteins provides stronger evidence for roles than individual pairwise relationships. Depending on the question of interest, individual, dyadic, or aggregate population analysis may be the most appropriate and effective.

Second, networks naturally represent heterogeneity among entities by virtue of local network topology. Rather than analysis on population aggregates (e.g. histograms), networks enable local querying of complex, non-metric attribute spaces. Non-metric properties of networks can be conceptualized through homophily, which *tends* to yield autocorrelation in attributes among entities close in the network. However, these correlations also tend to be non-monotonic: the *most similar* node to a query entity may be arbitrarily distant in the network. Due to this autocorrelation, network topology often represents local subspace clusters as overlapping, heterogeneous relationships. For example, a user's 'friends' in an online social network often clusters into functional units: friends from work, school, from the user's hometown etc. where each cluster are correlated in some–often unknown–attribute (e.g. my cycling friends). The effectiveness of simple heuristics such as counting common neighbors in the link prediction problem [Liben-Nowell and Kleinberg 2007] demonstrates the latent local information within social networks.

Third, networks are *interpretable* models for further analysis and hypothesis generation. Researchers can visualize small networks and examine relationships between nodes to compare against their knowledge and intuition in the domain. Furthermore, descriptive network measures enable researchers to compare networks according to density, degree distribution, clustering coefficient, centralities, diameter, average path length, triangle counts [Itai and Rodeh 1978; Tsourakakis et al. 2009] and graphlet distributions [Pržulj et al. 2004]. Many higher-level network measures have also been developed including robustness [Chan, Tong, and Akoglu Chan et al.; Purohit et al. 2014] local information efficiency [Babaei et al. 2016], and routing efficiency [Leskovec and Horvitz 2014; Watts and Strogatz 1998]. Using these shared measures, researchers can reason about the network through this comparison.

Finally, networks are *common* models for data, and can be re-used in multiple studies. The breadth of tools and support for network analysis allows researchers of various disciplines to apply sophisticated off-the-shelf analysis and visualization techniques, as well as easier storage, querying and portability in off-the-shelf graph databases. Finally, researchers have a common vernacular and skill-set developed in the area of "network science," despite originating from various domains such as biology or physics.

When inferring networks from non-network data, researchers ought consider whether higher-order structures are meaningful and informative, and which of these descriptive measures are appropriate on the inferred network. Arguably, available tools and convenience can motivate researchers to translate their problem into a network formulation, whether or not a network is the best model for the question of interest.

1.2. Inferring Networks from Non-Network Data

We define *inferred networks* as a class of network where the node and/or edge definitions are inferred from non-explicitly relational data. Work in machine learning and network mining typically focuses on applications of *explicit networks*: network representations where the meaning of nodes and edges is unambiguous and categorical, or weighted with low uncertainty. For example, in the (explicit) Facebook network, two adjacent users are categorically "friends." In weighted networks such as a road transportation network, nodes represent road intersections, and edges the road segments between them. These edges change with low probability (e.g. repair and construction) and are unambiguous. Measuring weights (e.g. travel time) is a matter of sensing/measuring traffic over the network, but these measurements are constrained to adjacent nodes on the known topology. In general estimating weights in these applications is a matter of accurately measuring a known relationship on a known topology, rather than learning a *hypothesized* relationship on an unknown topology.

Researchers routinely construct networks from attributes or other data to evaluate against their explicit network. Furthermore, much of the work in scientific domains (including biology, ecology, chemistry) learn *interaction* networks in the absence of ground truth, and have developed varied strategies to evaluate the quality of networks for prediction, classification, or the discovery of new relationships unknown in the domain.

1.3. Challenges

Inferring networks from non-network data provide several unique challenges:

- The relationships of interest within real datasets are often noisy, and confounded by overlapping relational structure at varying scales (e.g. temporal, spatial).
- Determining whether a particular method accurately encodes the relationship of interest in the network requires: (1) ground truth data, (2) model assumptions (e.g. Exponential Random Graph Model), or (3) some stability assumption (e.g. predictability over time). In many instances, no such assumptions or data are available, and researchers are left to tuning an interaction threshold.
- Many network topology inference methods with thresholds or other parameters define a continuous space of possible networks, often leading to ad-hoc selection or parameter-space sampling (see: Section 2.5.1). Although sensitivity analysis is routine for the model parameters of the subsequent task on the network (e.g. prediction, classification), most predictive models do not allow incorporating the inferred network topology into the model.
- The validity of descriptive interpretation of edges, paths, and modules is increasingly challenging with model complexity. For example, a social network defined via thresholding on who-calls-whom mobile call record data is more interpretable than a linear regression model on a feature vectors of mobile users. Determining the validity and interpretability of these higher-order features is crucial to the many analyses that use them.

1.4. Contributions

This survey organizes recent work focusing on inferring networks from non-network data in the construction of inferred networks, drawing from several domains where these methods are being applied as well as general methods in machine learning. This survey has two distinct contributions: (1) this work proposes a data science-focused formal description for the network topology inference problem, unifying work across

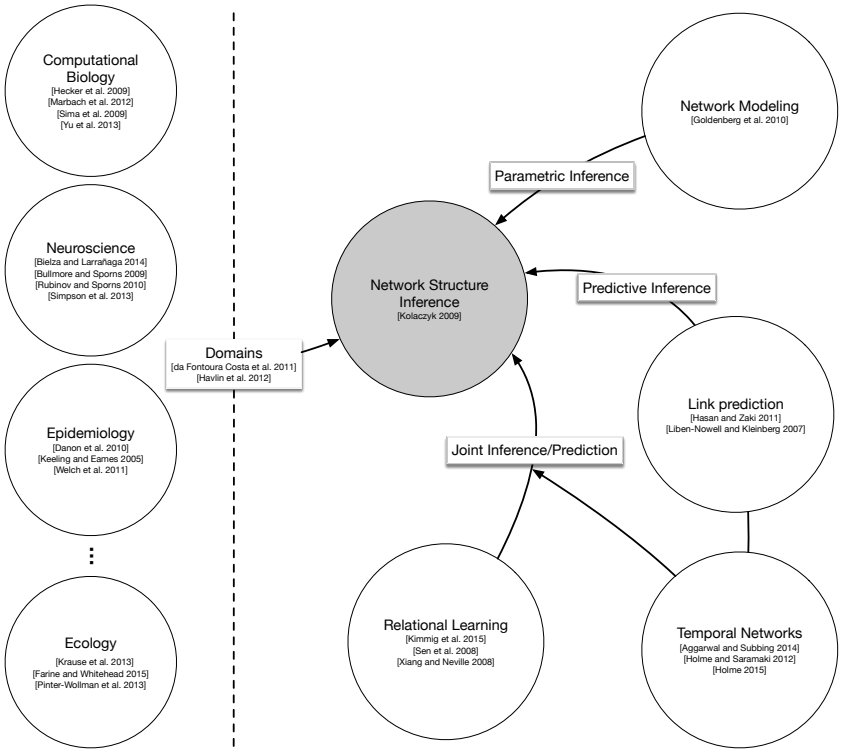


Fig. 1: An overview of related areas in machine learning and network science for this survey, several surveyed domains, and principle surveys and introductory work for these sub-areas.

many domains, and (2) we provide a taxonomy and shared vocabulary to organize the problem space in order to direct future research under a shared problem description.

1.5. Meta-Review: Comparison to Existing Surveys

In this section, we provide a brief meta-review of existing surveys and work which is related but distinct from the Network Structure Inference Problem. See Section 2.3 for how these areas relate to our proposed taxonomy of related work within this problem.

Figure 1 provides a map of research in related but distinct network science problems. The most similar work to this survey is [Kolaczyk 2009]. The author organizes related work in three different categories: (1) the link prediction problem, where some edges and all nodes are known and the task is to infer new edges, (2) interaction networks where all nodes are known, and the task is to infer edge relationships (e.g. by correlation), and (3) network tomography, where some edges and nodes are known, and the task is to infer 'interior' (unobserved) node and edge topology [Haddadi et al. 2008; Ni et al. 2010; Zhou et al. 2011]. Of these, (2) is primarily within the scope of what we define as the network inference problem; we focus with greater depth on the network inference procedure where no edge definition is known a-priori and must be learned.

We define the 'network structure inference problem' distinct from the large body of work in relational learning.[1] One branch within this area is attribute inference

[1]Within this area, our problem is broadly referred to as 'information extraction.'

and prediction *on* (explicit) networks. Given a network, these methods infer missing attributes using local estimates [Sen et al. 2008; Xiang and Neville 2008], or predict edges at a later time-step or by node attribute similarity [Hasan and Zaki 2011; Liben-Nowell and Kleinberg 2007; Lü and Zhou 2011]. Link prediction is one particular task of our general task-oriented framework.

A second branch in this area focuses on inferring probabilistic relational models from data [Getoor et al. 2007; Kimmig et al. 2014]. While graphical models are one strategy applied to our problem, generally these models treat attributes or variables as entities and are suited for semi-structured, often transactional data. Previous work in probabilistic relational models have learned relationships between an *explicit* input graph, and a learned output graph using node attribute inference, entity resolution, and link prediction tasks in a *supervised* setting [Namata et al. 2015].

Our survey draws on several application areas, however, we focus on comparing methodologies and challenges in these areas. da Fontoura Costa et al. [2011]; Havlin et al. [2012] span a more exhaustive range of application domains and their study of complex networks. Other recent surveys cover broad statistical network modeling [Goldenberg et al. 2010] and multilayer networks [Boccaletti et al. 2014; Kivelä et al. 2014]. Our survey draws on parametric network models, which is one class of inference method trying to infer the model parameters. Research in network fusion on multi-layer networks for a particular task (e.g. prediction) is one case of network structure inference where our input data is also relational. Recent surveys also outline work on temporal networks [Aggarwal and Subbian 2014; Holme 2015; Holme and Saramäki 2012]. Many network structure inference applications define edges by association measures over time (e.g. correlation in time series), so dynamics are an important aspect of network models for prediction tasks. These surveys cover each of these complimentary aspects in greater depth, and typically focus on explicit networks.

Our survey covers several different application areas. Recent domain-focused meta-studies [Marbach et al. 2012] and surveys in computational biology [Hecker et al. 2009; Li et al. 2008; Sima et al. 2009; Yu et al. 2013], ecology [Farine and Whitehead 2015; Krause et al. 2013; Pinter-Wollman et al. 2013; Proulx et al. 2005], neuroscience [Bielza and Larrañaga 2014; Bullmore and Sporns 2009; Rubinov and Sporns 2010; Simpson et al. 2013], political science [Lazer 2011] and epidemiology [Danon et al. 2011; Keeling and Eames 2005; Welch et al. 2011] all have significant discussion of network topology inference specific to the domain. However, none of these have network structure inference as a methodological focus and are limited to discussion of the single domain. Our survey focuses on challenges *across* each of these areas. We provide value to domain researchers both within and across fields, as well as researchers in machine learning interested in model development on networks.

2. PROBLEM DESCRIPTION

2.1. Preliminaries

We define a network $G = \langle \mathbf{V}, \mathbf{E}, \mathbf{A} \rangle$ as a tuple containing a set $\mathbf{V}$ of n nodes, $|\mathbf{V}| = n$, a set $\mathbf{E}$ of m node pairs $e_{ij} \in \mathbf{E}, |\mathbf{E}| = m$, and set $\mathbf{A}$ containing node or edge attribute sets. A particular attribute, the weight of an edge w_{ij} is a scalar value, $|w_{ij}| \leq 1$, where $w_{ij} = 0$ denotes the absence of an edge. An unweighted network is a special case of a weighted network where $w_{ij} \in \{0, 1\}$. Edge and node *features* are a particular type of attribute, derived by a kernel function measuring some local edge or node property (e.g. node degree). Time-varying network definitions are simply a t-length sequence of static network snapshots: $\mathbf{G} = (\mathbf{G}_1, ...\mathbf{G}_k, ...\mathbf{G}_t)$.

Symbol	*Definition*
$\mathbf{D}$	A dataset for input to a network inference method,(temporal, spatial, multivariate etc.)
$\mathbf{G}, \mathbf{E}, \mathbf{V}, \mathbf{A}$	$\mathbf{G} = \langle \mathbf{V}, \mathbf{E}, \mathbf{A} \rangle$, an attributed network with nodes $\mathbf{V}$, edges $\mathbf{E}$, and attribute-set $\mathbf{A}$
n, m	n number of nodes and m number of edges on network G.
v_i, e_{ij}, w_{ij}	node $v_i \in \mathcal{V}$, edge $e_{ij} \in \mathbf{E}$, edge weight w_{ij}
$\mathcal{R}(), \mathcal{T}()$	A network data model function $\mathcal{R}(\mathbf{D}, \bullet) \rightarrow \mathbf{G}$ producing graph $\mathbf{G}$ from input data $\mathbf{D}$, a task model $\mathcal{T}(\mathbf{G}, \bullet) \rightarrow (p_1, p_2 ...)$ producing a stream of responses from input graph $\mathbf{G}$.
$e()$	An error function $e()$ evaluating a task $\mathcal{T}()$.
τ	Edge similarity threshold for inclusion in the inferred network, $\tau \leq w_{ij} \Leftrightarrow e_{ij} \in E$

Table I: Table of symbol definitions used throughout this survey.

2.2. Data Science Motivations for Network Structure Inference

Data science has been a growing and somewhat contested[2] field of study in recent years. Under the name 'Data-intensive science,' Jim Gray argues that data and computation enable a new 'fourth' paradigm of science. These data-driven models enable the investigation of questions beyond the *capability* of science utilizing theoretical model-driven (the second) or model-driven simulation (the third) paradigms [Hey et al. 2009]. The scope of what can be called 'data science' is heavily overloaded, ranging from high-performance computing and computing at scale, to visualization, to blog posts analyzing political census data or networks of characters in film scripts. 'Data science' in some of these numerous contexts arguably inverts the traditional process of science, with a strong focus on exploratory analysis rather than hypothesis testing and experimental design. In this exploratory setting, data scientists discover surprising relationships and hidden business value in large, complex datasets, or focus on predictive modeling without a motivating question on the underlying system.

We organize the network inference problem through a lens of hypothesis-driven data science. Under this perspective, the value of a network can be stated simply: for a scientific question of interest, and its relevant (non-network) data, are networks an informative and useful *data model* for *better* answering our question?

Figure 2 gives an overview of this data science workflow, and all the relevant terms to our taxonomy. As researchers, we start with a broad question of interest, and try to locate or collect the data relevant to answering our question. There are several possible data sources and modalities for analysis, coupled with several possible representations (including networks). Our choice of data representation informs and constrains our hypotheses about the question of interest, and typically domain science (orange, top) and machine learning (purple, bottom) generate complimentary results for hypothesis generation. Novel computational models are developed to test these hypotheses. Results from this data-driven paradigm have closed the gap in understanding for many questions in novel ways. Data-driven science has also developed novel methodologies, enabling new, large-scale experimental design [Backstrom and Kleinberg 2011; Gui et al. 2015] and randomization techniques [Efron and Tibshirani 1993; Kleiner et al. 2014], two key methodologies across scientific domains.

Figure 2 pinpoints several levels of modeling which can impact the final performance at answering our question of interest. Often different teams will be responsible for generating and collecting the underlying data, inferring or defining the network, or developing the predictive models. For example, machine learning researchers will rarely control the underlying sampling rate of very specialized data collection workflows in

[2]With several disciplines claiming ownership or criticism.

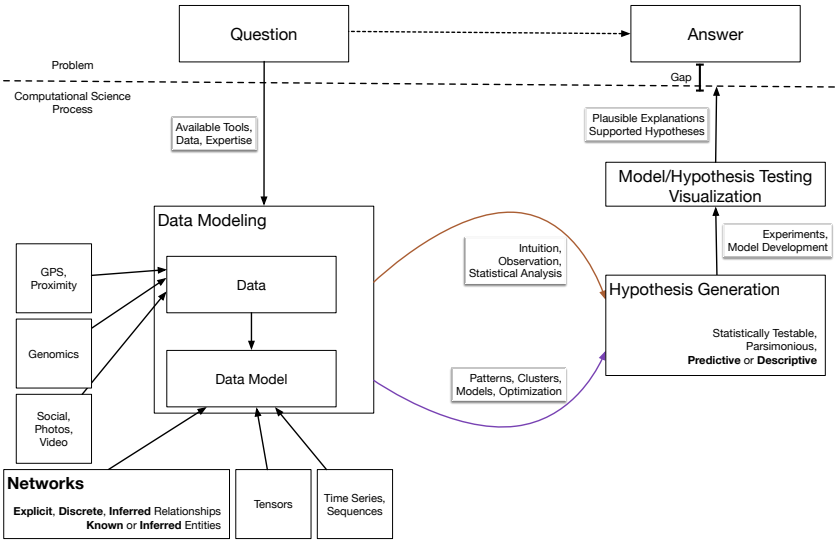


Fig. 2: A general framework for computational science and data science, with a focus on networks as data models for non-network data. A researcher typically (1) uses *relevant* data and *appropriate* data models to formalize a particular question. Traditional domain science (top, orange) typically combines direct observation and simple statistical models. Data-driven science with machine learning (bottom, purple) augments this intuition, hypotheses and statistical models with sophisticated features and patterns, yielding additional challenges for visualization, hypothesis testing and generation. The output of this analysis is a plausible explanation or some supported hypothesis. The result of data-driven science are considered to be shrinking this 'gap' in understanding, toward robust and verifiable answers to the original question, in novel ways (see: [Hey et al. 2009]).

bioinformatics, geophysics and climate, or neuroscience. In the machine learning community, there is a great deal of sensitivity analysis for the subsequent predictive model parameters, without testing the parameters of the underlying network representation. This strategy is effective at producing sophisticated models "given a network," but is less suited at answering our scientific questions on the original data.

2.3. A Taxonomy for Network Structure Inference

We taxonomize work in the network inference problem according to (1) the varying 'difficulty' of defining edges and nodes given the nature of the *underlying data* of the network, (2) the *type of method* (e.g. regression, correlation, novel interaction measures) used for inference, and (3) the types of *questions* and hypotheses which are studied using the network representation.

2.3.1. Explicit, Discrete, and Inferred Edge Definitions. Networks model three broad types of relationships (i.e. edges), typically based on a measure of interaction between entities. Categorical, *explicit* relationships are unambiguously known in the system–such as the 'friend' relation in Facebook. *Discrete* interactions denote unambiguous transactions occurring between two entities–such as phone calls or text messages in mobile device data. The primary task of defining edges on these interactions is to select an appropriate threshold to measure the strength of the relationship. *Inferred* interactions denote some statistical measure of similarity, beyond simple transaction counts.

For example, the definition of a spatiotemporal *co-location* interaction between two entities can be simply inferred by specifying "how close" for "what duration." This is generally a more challenging measure of relationships relative to discrete interactions.

2.3.2. Explicit and Inferred Node Definitions. Similarly, inferring nodes is of varying difficulty in different domains. In most applications, the nodes are explicit in two ways: (1) the node definition is unambiguous (e.g. a 'user', 'animal' or 'gene' is a node), and (2) there is a node *correspondence* between time-steps in the data (e.g. this particular node corresponds to user 'Ivan' at each time step). Functional brain networks, and climate networks are two examples of domains where the node definition is not explicitly given–e.g. a group of two-dimensional earth surface pixels, or three-dimensional brain voxels. The *scale* of node definition can dramatically change the performance of predictive models, or the stability of descriptive statistics derived from the network [Cammoun et al. 2012].

2.3.3. Hypotheses and Tasks on Inferred Networks. Before discussing data models, we should first have some intuition of what a researcher wants to *do* with the network. Figure 2 (right) illustrates the core novelty in machine learning and data science in hypothesis generation, model development, and testing. In machine learning, there is typically a bias toward evaluating these hypotheses using a novel prediction task model. However, typically these models do little to explain the underlying system, and are more useful for generating new hypotheses. In the statistics domain, there is a bias toward descriptive modeling, where the 'task' is parameter inference on an assumed parametric (or "non-parametric," [Wasserman 2006]) statistical model on the data. These parameters are useful at interpreting the underlying system, such as fitting the exponent of degree distributions, or reporting other aggregate statistics on a network.

Applications driven by descriptive models aim to reconstruct and describe the underlying relational structure with the greatest fidelity, relative to the domain knowledge base. A typical example in this area are gene regulation networks (GRNs). These networks are inferred on data measuring individual genetic expression over different experimental settings. In these networks, nodes represent genes or functional gene families, and edges represent inferred positive or negative gene expression relationships (e.g. "gene 'A' reduces the expression of gene 'B' under some context"). Inference of these networks typically identifies new, high quality *candidate* regulation relationships given high accuracy of inferring known edges. These interactions can be experimentally tested to build greater understanding of cellular processes and to develop potential personalized medical treatments.

In contrast, predictive methods aim to discover topology within the network which maximizes predictive performance, and may not reconstruct the underlying process with the highest fidelity, but with a focus on those aspects or modalities which are most predictive. Modeling the predictive aspects of the data allows researchers to learn regular relationships between modalities (e.g. call, SMS, and location in mobile phone data) or over time (e.g. periodicities) and aids further hypothesis generation to explain the predictive relationships. However, often highly predictive relationships are also uninteresting and can drive the structure of the inferred network. Domains such as climate tend to subtract known periodic dynamics as preprocessing on the underlying data, prior to inferring the network structure.

2.3.4. Methods for Network Structure Inference. We organize related work broadly along the type of network structure inference *model* used, including parametric, non-parametric and various thresholded interaction/correlation measures. Within these

groupings, we categorize the type of task performed on the network, including edge and attribute prediction, descriptive analysis, or model selection.

Table II summarizes work across several domains, introducing the basic scientific question driving the analysis, and the network structure inference model used to realize the network. In Table II (Column 'Model'), we label these models under two broad categories. First, *parametric models* allow interpretable, descriptive statistics. We identify graphical models (**GM**) and other network models fit with maximum likelihood methods (**ML**), relative to some assumption on the input data such as information flow between nodes [Gomez-Rodriguez et al. 2012]. Causal models (**CM**) typically generate causal networks–a special case of graphical models–using Granger causality [Granger 1969] or other causal definitions [Mani and Cooper 2004; Meek 1995]. These networks represent strong relationships between nodes which control for confounding factors caused by other (possible) adjacent nodes.

Second, *non-parametric models* tend to directly measure interactions between nodes and use statistical tests to determine appropriate edge weights. Section 2.5.1 covers this category in greater detail. We categorize work related to novel and 'ad-hoc' interaction measures (**I**) between the data associated with pairs of nodes, correlation networks (**IC**) which measure linear, cross, or some other correlation, entropy (**IE**), frequency domain measures (**IF**), and regression (**R**).

Table II (Column 'Task') categorizes rows within each domain by the type of task performed, under the caveat that one study may use several evaluation strategies, or that the actual task could only loosely be described as the canonical task (e.g. edge prediction).

First, we denote predictive tasks, including edge prediction (**PE**) and attribute prediction (**PA**). Attribute prediction can also describe prediction of the original data, i.e. by simulating/generating data through the network model [Papalexakis et al. 2014]. Predictive models are relatively rare across domains because researchers are very interested in expressive models which give insight into the underlying system. We observe some specialization in both information networks (in the machine learning literature) and epidemology, which aims to predict the extent or timing of an epidemic over a population in varying contact models.

Second, descriptive analysis is broken into node-oriented statistics (**DN**), this includes reporting distributions of simple node statistics including degree distribution, clustering coefficient, correlation distributions, etc. This often constitutes the base-level exploratory analysis. Role-oriented analysis (**DR**) aims to characterize nodes using network features, by the structural roles they play in the system (i.e. bridges in-between social communities). Other high-order analysis (**DH**) examines communities or other larger subgraph structures beyond node and edge-based descriptive statistics.

In both descriptive and predictive cases, we observe a good deal of work in model selection (**MS**). These varying models correspond to different hypotheses to how the network might have been generated.

2.3.5. Evaluating Networks and Tasks. Evaluation strategies over the entire workflow of Figure 2 measure the performance of the *final task* (prediction or statistical inference). In this context, the network serves as a model of the data, but the fidelity of this model in terms of fitting error, or evaluation against partial ground truth network data does not measure the network's usefulness at answering questions on the original input data. For any non-trivial data or domain, it is perhaps more appropriate to think of the space of networks as *possible* representations with some utility for answering a specific question. 'The network' is typically seen as uncovering the true relational structure of the data with some error [Wang et al. 2012].

Most scientific domains agree with an evaluation strategy focused on final task (a network *for* location prediction, brain activity response etc.), because there is no evaluation network for comparing the inferred network structure. For example, we simply cannot survey baboons [Farine et al. 2016] as we can humans [Eagle et al. 2009] to discover their real friendships. Furthermore, uncovering the general, robust network from complex data may not appropriately model the overlapping modalities of the data. For example, a multitude of complex contexts might mean that functional brain networks are inherently probabilistic. Discovering the underlying, robust network model will be less informative in the general case because it does not account for this complexity.

2.4. Problem Definition

The Network Topology Inference problem represents some input data as a network and *validates* this network relative to performance on some task(s). This is an active area of research across many application domains, but lacks a general framework for evaluation, typically using domain rules-of-thumb and application-specific evaluation schemes. This is an initial work formalizing this problem, particularly in the absence of ground-truth network data.

The network topology inference problem combines two basic models: first, a *network model* $\mathcal{R}(\mathbf{D}, \bullet) \rightarrow \mathbf{G}$ constructs network $\mathbf{G}$ on input data $\mathbf{D}$. This model may be parametric statistical models (e.g. ERGMs) or non-parametric and threshold-based interaction networks. Second, the problem uses a *task model* $\mathcal{T}(\mathbf{G}, \bullet) \rightarrow (p_1, p_2 ...)$ on input $\mathbf{G}$ under some parameters, which emits task responses (e.g. prediction 'p_i'). These responses approximate the hidden, ideal function $\mathcal{T}^*(G)$ of a particular network task (e.g. classification, prediction) with error $e()$.

This formulation may seem underspecified. However, it succinctly clarifies the relationships between data input, the network model, and the task model. It explicitly formulates network $\mathbf{G}$ as a *model* on data $\mathbf{D}$ for task $\mathcal{T}^*$, approximated by $\mathcal{T}$. This formulation captures simple *interaction network* workflows (see: Section 2.5.1) which separately infer the network (often by expertly-tuned thresholds) and validate task performance, as well as parametric inference methods which learn the network model parameters (and possibly, jointly the task model parameters). To our knowledge, all network inference models can be formulated in this pattern, and all network inference models *should* be formulated relative to a particular task or hypothesis. In much of the existing work, the network model or task model will not be formulated explicitly, or the space of possible model combinations may be under-explored.

We can instantiate several tasks within this framework. In the context of network prediction tasks, our predictive model can output predictions of (1) edges, (2) attributes, or (3) the original data. For one instantiation, on a validation edge-set $\mathbf{E}^*$, we can evaluate:

$$\underset{\mathbf{G}}{\operatorname{argmin}}\, e(\mathcal{T}(\mathcal{R}(\mathbf{D}, \alpha), \beta), \mathbf{E}^*) \tag{1}$$

This is a joint optimization of $\mathbf{G}$ over parameters α and β, for edge prediction. In this context, the suitability of both the network and task models (and the appropriate error function) will determine the performance of the inferred network. This joint model may an EM-like formulation, where the task and network models share information iteratively to re-learn their parameters.

This formulation also demonstrates two key challenges in the network inference problem. First, the parameter space of possible $\mathbf{G}$ from $\mathcal{R}(\mathbf{D}, \bullet)$ may be large, and the joint optimization with the parameters of a given task is likely to be non-convex. Furthermore, well-performing local solutions in the task parameter-space may have

different network topologies. Summarizing and reconciling these differences may be important in interpreting where the network is most robust. It is unclear how to reconcile differences and summarize this space of *plausible* networks into a single, interpretable, effective model. This proliferation of possible task and network model combinations makes further hypothesis generation and testing daunting for the researcher, who has diminishing confidence in empirical statements about the system of interest.

2.5. Formulating Data-to-Task Workflows with Network Data Models

Our problem definition is flexible enough to incorporate the varying types of network models (e.g. regression, parametric network models) and tasks (e.g. prediction, descriptive statistics and hypothesis testing). Below, we describe several formulations for network structure inference, according to their network model and task model.

2.5.1. Interaction Networks. The most prevalent class of network topology inference is measuring pairwise interactions (e.g. correlations) between nodes, and choosing a threshold to define a sufficient degree of interaction. This threshold may be chosen by some statistical test, by tuning on some desired criteria (e.g. assumed network density), or in an ad-hoc way with prior domain knowledge.

This has been discussed in the context of *discrete* interactions as described above:

> "Inferring networks from pairwise interactions of cell-phone call or email records simply reduces down to selecting the right threshold τ such that an edge (u, v) is included in the network if u and v interacted more than τ times in the dataset. Similarly, inferring networks of interactions between proteins in a cell usually reduces to determining the right threshold." [Myers and Leskovec 2010]

Researchers often make several application-specific decisions around these thresholds:

> "From this complete correlation graph, only the edges with significant correlation (> 0.5) were retained. But using the same threshold for positive and negative correlations is not appropriate as negative correlations are usually weaker and many nearby locations have high positive correlation" [Kawale et al. 2013]

> "We let δ as a user-controlled parameter, where larger δ values correspond to less predicted regulations, and only focus on designing a significance score s(t,g) that leads to 'good' prediction for some values of δ" [Haury et al. 2012]

These methods typically produce a fixed network model but do not explore the space of possible networks under varying interaction thresholds, except through offline trial-and-error. Furthermore, subsequent sensitivity analysis on these outputted networks G are often performed on the task model parameters alone, rather than jointly on both models.

We can formulate the workflow for this analysis under our framework. Assume the interaction threshold τ is given by hand-tuning or domain knowledge, we have some feature matrix $\mathbf{D}$ that has some notion of similarity between features, and $\mathcal{T}()$ is an edge prediction task evaluated on $\mathbf{E}^*$. Then this workflow is expressed as:

$$\mathcal{R}(\mathbf{D}, \tau) \rightarrow \mathbf{G}; \underset{\beta}{\operatorname{argmin}}\, e(\mathcal{T}(\mathbf{G}, \beta), \mathbf{E}^*) \tag{2}$$

When these interaction networks are evaluated in the absence of ground truth, the network may be measured through autocorrelation. In this case, the same network

inference is applied, $\mathcal{R}(\mathbf{D}^*, \alpha) \rightarrow \mathbf{G}^*$ for some hold-out data D^* (e.g. at some later time).

2.5.2. Parametric Inference and Model-Fit Networks. Maximum-likelihood methods assume some parametric model family to represent relationships between nodes (such as time between interaction, likelihood of information transmission over time) and infer the best model parameters. For clarity, we work through one specific application, in epidemiology and information networks such as blogs, although the pattern is similar in other applications. Structure inference methods on information networks share the assumption that we are observing the 'arrival' of information or attribute value at nodes (i.e. computers, blog pages, individuals) over time, but are unable to observe the topology which transmitted the information. For an edge e_{ij}, the infection time difference $t_j - t_i$ can be fit against an infection model over time, measuring the likelihood that v_i infected v_j [Myers and Leskovec 2010].

Where input data $\mathbf{D}$ are infection times of each node, we can formulate these workflows as:

$$\underset{\alpha}{\operatorname{argmin}}\, \mathcal{R}(\mathbf{D}, \alpha) \rightarrow \mathbf{G}; \underset{\beta}{\operatorname{argmin}}\, e(\mathcal{T}(\mathbf{G}, \beta), \mathbf{E}^*) \tag{3}$$

In information network applications, $\mathbf{E}^*$ is typically provided by a known network. Processes are simulated on this network to generate input data for the maximum-likelihood relational data model $\mathcal{R}()$. This method is used to 'reconstruct' $\mathbf{E}^*$ only from input data $\mathbf{D}$.

2.5.3. Jointly-Learned Network and Task Models. Previous work in statistical relational learning on *explicit* networks has focused on jointly learning relationships between (categorical) attributes and a predictive task, such as link prediction [Gong et al. 2014; Namata et al. 2015], and distinguishing correlated effects between these processes [La Fond and Neville 2010].

Previous work has also used a *workflow* approach to the network structure inference problem which maximizes performance of particular task(s) on the inferred network [De Choudhury et al. 2010; Farine et al. 2016]. Consider interaction networks over varying thresholds τ. A naive solution for this type of approach is to explore the parameter space of $\mathcal{R}(\mathbf{D}, \tau)$ and evaluate the task performance on each inferred network. However, this is very costly, especially as $\mathcal{R}$ requires more parameters.

The joint optimization for network structure inference learns the features in $\mathbf{D}$ that perform well at task $\mathcal{T}()$. For example, previous work has learned network models which jointly discovers features in a supervised LDA model, and the logistic regression weights which perform well for an edge prediction task [McAuley et al. 2015]. In Equation 1, we have already formulated joint optimization network structure inference within our framework, repeated here:

$$\underset{\mathbf{G}}{\operatorname{argmin}}\, e(\mathcal{T}(\mathcal{R}(\mathbf{D}, \alpha), \beta), \mathbf{E}^*) \tag{4}$$

The jointly optimized model may avoid discovering spurious relational model parameters which are not suitable for the intended task. Second, we may be able to learn a more interpretable relationship between the original feature space and the task model. However, this strategy has the added requirement that the output of each model can be used to re-train the other. For example, McAuley et al. [2015] use a *supervised* LDA model which uses the output of the edge prediction task to re-train the relational model. Because of this added modeling complexity, joint models are relatively rare compared with multi-step workflows.

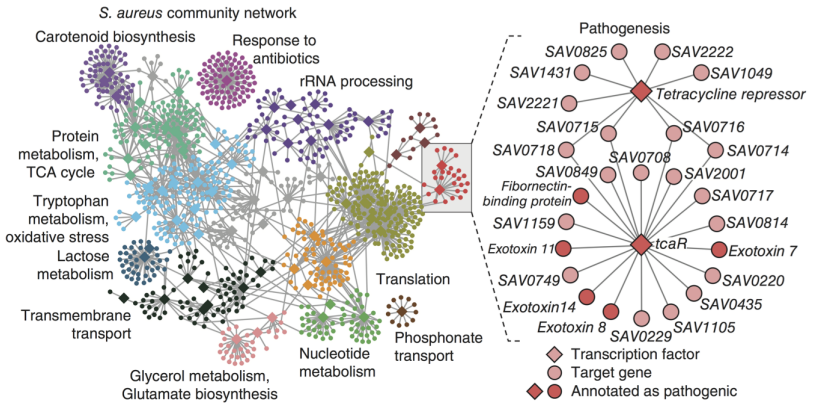


Fig. 3: From [Marbach et al. 2012]. A consensus gene regulatory network (GRN) of *Staphylococcus aureus* generated from microarray data in the DREAM5 challenge ensemble workflow (see: Figure 4). Network modules in agreement with Gene Ontology (GO) database are colored and labeled according to their shared function (grey are not coherent in GO). The detail (right) shows inferred gene regulation related to pathogenesis.

3. APPLICATIONS OF NETWORK STRUCTURE INFERENCE

Although the applications in 'network science' are far too broad to survey with any meaningful focus, we look closely at a few applications where inferring networks from non-network data is a particular focus. Alongside these inferred networks, several of these applications have extensive work on networks from direct observation (e.g. in ecology).[3]

3.1. Computational Biology: Discovering New Genetic Regulation

Networks are constructed in computational biology to model many different processes. Protein-protein interaction (PPI) is the most common type of network in this domain, constructed experimentally through yeast two-hybrid experiments which physically test for binding of one protein to another. Metabolic networks are a process network which models relationships between enzymes, metabolites (nodes) on processes (edges) such as reactions or pathways (e.g. biologically meaningful paths through the network). We focus only on gene regulation networks (GRNs) and gene co-expression networks (GCNs) which have focus on statistical inference of relationships between genes under different experimental scenarios (for further reading, see: [Sima et al. 2009]).

3.1.1. Underlying Data in Gene Regulatory Networks. Next-generation high-throughput microarray technologies allow the sequencing of genomes and measuring the *expression* of particular genes at a large scale and low cost [Shendure and Ji 2008]. 'Expression' is the measurement of how groups of genes produce different phenotypic specializations through the production of different proteins. In cellular development, different gene co-expression can be responsible for RNA translation or nucleotide metabolism, yielding many complex functions from gene interaction (see: [Barrett et al. 2013]). Figure 3 illustrates a small network with annotated functional clusters.

[3]All figures reprinted with permission and attribution.

Domain	*Problem*	*Sec.*	*Model*	*Task*	*Citations*
Computational Biology	Discover interactions between genes in cellular processes	3.1	**I**	**DH**, **DR**	Zhang and Horvath [2005]
			IE		Butte and Kohane [2000]; Faith et al. [2007]; Meyer et al. [2007]
			GM		Allen and Liu [2012]; Lebre et al. [2010]; Mani and Cooper [2004]; Mukherjee and Speed [2008]; Toh and Horimoto [2002]
			R		Haury et al. [2012]; Yuan and Lin [2006]
			CM	**PE**	Barzel and Barabasi [2013]; Feizi et al. [2013]
Climate	Describe relationships in environ-mental system dynamics	3.2	**IC**	**DN**, **PA**	Kawale et al. [2012]; Paluš et al. [2011]; Steinhaeuser et al. [2011]; Tsonis and Roebber [2004]; Yamasaki et al. [2008]
			IE		Donges et al. [2009b]; Hlinka et al. [2013]
			CM	**DN**	Ebert-Uphoff and Deng [2012]; Kretschmer et al. [2016]; Runge et al. [2013]
			GM		Zerenner et al. [2014]
			R		Zhou et al. [2015]
Neuroscience	Model relationships between brain regions, physiological structure, and function	3.3	**IC**	**DN**, **MS**	Bialonski et al. [2011]; Zalesky et al. [2012]
			IF		Lachaux et al. [2002]; Pfurtscheller and Andrew [1999]; Ponten et al. [2016]; Zhan et al. [2006]
			CM		David et al. [2008]; Dhamala et al. [2008]; Friston et al. [2011]; Ramsey et al. [2010]; Roebroeck et al. [2005]; Rosa et al. [2012]
			ML	**PA**	Papalexakis et al. [2014]
Epidemiology	Model hidden networks from observed infections	3.4	**I**	**PA**, **MS**	Adar and Adamic [2005]; Haydon et al. [2003]
			GM		Britton and O'Neill [2002]; Groendyke et al. [2011]; Stack et al. [2013]
			ML		Du et al. [2012]; Gomez-Rodriguez et al. [2014, 2012]; Myers and Leskovec [2010]; Netrapalli and Sanghavi [2012]
Ecology	Describe and predict animal behavior	3.5	**I**	**DH**, **MS**	Aplin et al. [2012]; Haddadi et al. [2011]; Psorakis et al. [2012]
			IE		Barrett et al. [2012]
			R	**DN**	Whitehead and James [2015]
Mobile	Predict social influence on individual mobility	3.6	**I**	**PE**, **DN**	De Choudhury et al. [2010]; Eagle et al. [2009]; Mastrandrea et al. [2015]; Sekara and Lehmann [2014]

Interaction, **IC**: Correlation, **IE**: Entropy, **IF**: Frequency, **I**: Novel measures
CM: Causal model **GM**: Graphical model, **ML**: Maximum likelihood, **R**: Regression

Prediction, **PA**: Attributes, **PE**: Edges
Descriptive Analysis, **DN**: Nodes, **DR**: Roles, **DH**: Other high-order, **MS**: Model Selection

Table II: A summary of related work, across domains

The output of the microarray analysis (with notable simplification) is a 2D data matrix of numeric values measuring the expression of a gene (row), on a particular experimental design, subject, or time step (column) [Bar-Joseph et al. 2012]. Defining edges between genes simplifies to comparing expression profiles across the different columns of the data.

3.1.2. Discovering New Gene Interactions From Data. The high-level 'task' for gene regulatory networks is *link prediction* on the network learned from data, to discover unknown gene regulation candidates which can be experimentally tested. The network model inferred from data should agree with databases of biologically-known interactions and function, while providing few verified false-positive regulations.

Table II illustrates that the inference of GRNs is very mature relative to other domains. gene regulatory network inference is very mature relative to other domains, since the inferred network has immediate value for future investigation and hypothesis generation, and it is verifiable according to current domain knowledge. Researchers in this area apply most categories of prediction task models including regression, correlation, mutual information, and graphical modeling. One unique challenge in gene regulation is the issue of confounding factors including indirect and transitive associations which lead to many spurious edges. Recent work has measured these 'direct' (e.g. causal) edges in noisy expression datasets [Barzel and Barabasi 2013; Feizi et al. 2013].

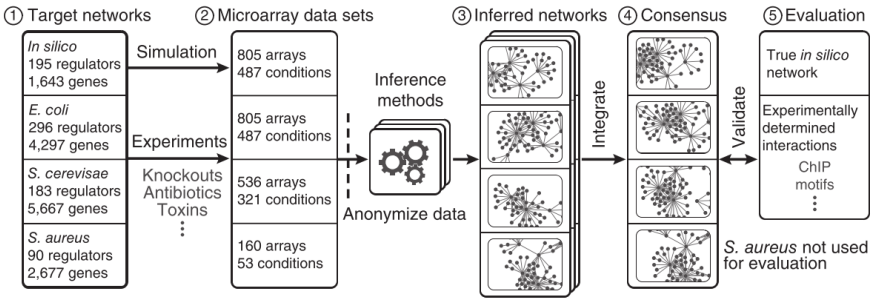


Fig. 4: From [Marbach et al. 2012]. The general workflow design for the DREAM5 network inference challenge. This workflow (1-2) generates one simulation dataset and three experimental microarray datasets from three well-studied species. The 29 participating inference methods all (3) generate inferred network output, (4) a consensus network is constructed for each dataset, and then (5) validated against known edges in synthetic networks, and against experimentally known edges in two of the real datasets.

3.1.3. A Workflow From Gene Expression Data to Interaction Discovery. Marbach et al. [2012] introduces an ensemble approach associated with the DREAM5 competition. The authors present 29 different network inference methods across different model types, including regression, mutual information, correlation (a.k.a. 'relevance networks' in this domain), Bayesian networks, ensembles, and other novel approaches (e.g. random forests, neural networks, Gaussian mixture models, etc.) This list demonstrates the maturity and variety of methods applied to this problem. These methods use pairwise (e.g. gene-to-gene correlation) or group-wise (e.g. many-to-one group LASSO) measures

of interaction intensity, which yield a directed, unweighted network signifying "gene A regulates gene B."

Figure 4 illustrates an ensemble workflow for the DREAM5 network structure inference challenge. The authors (1) generate a ground truth network on three species using experimental trials as well as synthetic network data. In (2), these experiments yield four different datasets of biologically-tested networks of gene regulation, as well as the raw (non-relational) gene expression. The collection of inference methods produce (3) inferred networks on each of the four datasets. The authors (4) integrate these 29 different inferred networks to produce an ensemble network. This network is (5) validated against the ground-truth networks generated in step (1). Finally, the authors show that the ensemble method discovered 59 potential interactions, of which 29 show some support, and 20 show strong support for being biologically significant.

We examine this process in such detail to demonstrate the 'complete' workflow for network structure inference, from input data, to relational model, to a final output (learning the regulatory network). Within our formulation, this workflow does not have an explicit *task model*, as the network itself is the object of interest. Therefore, this application is typically *descriptive* modeling against a known ground truth. Although these models do 'predict' new potential gene regulation via previously unknown edges, these interactions are usually checked manually through experimentation.

3.2. Environmental Science: Discovering New Climate Relationships and Predicting Outcomes

Networks inferred to understand climate dynamics are among the most difficult to model of any domain, and much of the work to formalize and validate these networks is actively being developed. Within this domain, researchers are interested in discovering robust, *causal* relationships between climatic variables, over different spatial regions of earth. This modeling can improve prediction of changing hydrological processes, land-cover, ecosystem productivity, and polar or sea ice cover, which are key aspects for climate change mitigation. Two unique challenges exist for inferring climate networks: (1) input data is typically noisy, highly spatially-autocorrelated, multivariate time series of climatic variables collected under varying regimes and sensor quality. Domain scientists produce "reanalysis" data products which attempt to mitigate these problems. However, inferring accurate networks from these data requires significant understanding of the data ingestion workflows [Levitus et al. 2013; Saha et al. 2014], and its introduced biases and variability. (2) the structure of climate networks is not well understood aside from a handful of climate indices–coarse spatial locations on earth where dynamics are well-studied and regulate or correlate with other environmental processes (e.g. El Niño and La Niña oscillation cycles). Therefore, validating correctness of the inferred network is suitable for unsupervised strategies such as relational or predictive modeling of the original data.

3.2.1. Network inference methods in climate and environmental science. Nearly all studies constructing climate networks use some time series similarity as an underlying relational measurement. Previous work has used linear correlation [Donges et al. 2009a; Steinhaeuser et al. 2011; Tsonis and Roebber 2004; Yamasaki et al. 2008] or mutual information [Donges et al. 2009a; Hlinka et al. 2013], and use either a hand-picked [Donges et al. 2009a; Tsonis and Roebber 2004] or simple statistical test [Yamasaki et al. 2008] to set similarity threshold τ–where similarity greater than τ is considered a binary edge in the network.

There is considerable focus on formulating these simple pairwise comparison methods, and often the 'recipe' of the network according to parameter settings and preprocessing choices varies greatly from study to study. These networks are typically binary

rather than weighted, because the final output of interest is a binary decision on the similarity distribution: (e.g. {"significant", "not significant"}). However, typically these measures will have no 'natural' threshold which gives this binary classification. Instead, these networks can be gradually densified or sparsified by loosening or tightening the similarity threshold.

Descriptive statistical work has been very popular *downstream* from these varying network 'recipes.' Donges et al. [2009a] reports clustering coefficient, betweenness centrality, closeness centrality. Tsonis et al. [2011] reports community structure which tends to cluster in spatially-contiguous locations, on account of the autocorrelation present in these networks. Little work focuses on evaluating these networks as predictive models for the input data. Steinhaeuser et al. [2011] use both descriptive statistics and predictive performance to evaluate the inferred network.

As work utilizing network models grows in this climate and environmental science, researchers have developed more sophisticated techniques for determining edge significance [Kawale et al. 2012], or conditional dependencies using causality [Ebert-Uphoff and Deng 2012; Kretschmer et al. 2016; Runge et al. 2013].

3.2.2. A Workflow from Environmental Sensing Data to Environmental Interaction Network. We will step through a concrete example of constructing a climate network from spatially-gridded time series data of global surface air temperature (SAT) [Donges et al. 2009a]. The authors measure the similar dynamics of pairwise earth locations (corresponding to nodes n_i, n_j) and construct edges between locations with 'significant' similarity. The authors use two standard measures, linear correlation and mutual information between time series X_i and X_j:

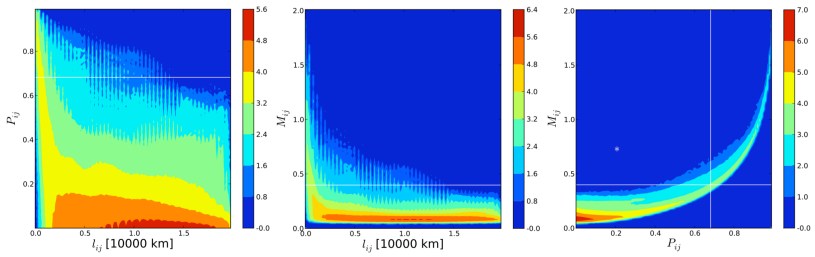
$$P_{ij} = \frac{\sum_{t=1}^{|X_i|}(X_{i,t} - \bar{X}_i)(X_{j,t} - \bar{X}_j)}{\sqrt{\sum_{t=1}^{|X_i|}(X_{i,t} - \bar{X}_i)^2}\sqrt{\sum_{t=1}^{|X_j|}(X_{j,t} - \bar{X}_j)^2}} \tag{5}$$

$$M_{ij} = \sum_{b=1}^{|B|} p_b(X_i, X_j) \log \frac{p_b(X_i, X_j)}{p_b(X_i)p_b(X_j)} \tag{6}$$

Equation 5 is the sample Pearson correlation between two time series, where $\bar{X}_i$ is the sample mean of time series X_i. The denominator represents the product of the sample standard deviations of X_i and X_j. This measures linear relationship of X_i and X_j over the length of the time series. Equation 6 is the discrete mutual information estimation between two time series, where $p_b(X_i, X_j)$ is the joint cumulative distribution of the b-th discretization window, and $p_b(X)$ the marginal cumulative distribution of the b-th discretization window. This measure compares the shape of the joint and marginal CDFs under the independence assumption. When the joint distribution is equal to the product of marginal distributions: $\log(1) = 0$ yields no mutual information (e.g. X_i and X_j are independent).

Varying similarity threshold τ produces networks of varying edge densities ρ and other network measures such as the size of the largest connected component. The authors select thresholds for correlation and mutual information ($\tau_{corr} = 0.682$ and $\tau_{MI} = 0.398$) such that they produce the same network density ($\rho = 0.005$).

Figure 5 (a) shows the density of pairwise linear correlation measures (P_{ij}) vs. geographic distance between nodes, on a logarithmic color bar scale. This illustrates a strong spatial autocorrelation between nearby points. (b) shows the pairwise distri-



(a) P_{ij} vs. distance (b) M_{ij} vs. distance (c) P_{ij} vs. M_{ij}

Fig. 5: From [Donges et al. 2009a]. On the input data of averaged global surface air temperature (SAT) at different spatial regions, (a) shows the distribution of pairwise linear correlation measures (P_{ij}) vs. geographic distance between nodes, on a logarithmic color bar scale. This illustrates a strong spatial autocorrelation between nearby points. (b) shows the pairwise distribution for mutual information pairwise calculation vs. geographic distance, showing less spatial autocorrelation. The horizontal bars indicate thresholds on the P_{ij} and M_{ij} scales which produce the same network density ($\rho = 0.005$). (c) shows the linear correlation vs. mutual information. The starred quadrant ((c) top-left) denotes edges defined by mutual information but not by correlation.

bution for mutual information pairwise calculation vs. geographic distance, showing less spatial autocorrelation. The horizontal bars indicate thresholds on the P_{ij} and M_{ij} scales which produce the same network density ($\rho = 0.005$). (c) shows the linear correlation vs. mutual information. The starred quadrant ((c) top-left) denotes edges defined by mutual information but not by correlation.

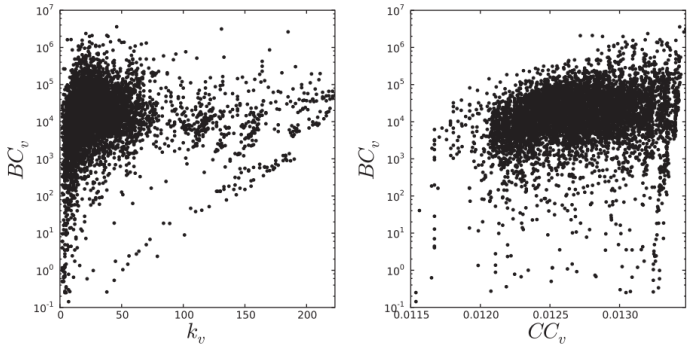
While several methods have been developed to test edge significance, little work has focused on the validity of higher-order structures such as paths or communities. While the graph *definition* of paths or communities are valid on these networks, no known work measures the *interpretation* of these relationships with respect to the original time series data, or domain knowledge.

Figure 6 explores the sub-spaces of different node measures on the inferred network. Figure 6(a) shows the distribution of betweenness centrality per node, vs. the degree per node. Similarly, 6(b) shows betweenness centrality and closeness centrality. The authors demonstrate that (degree-preserving) edge re-wiring randomization indeed destroys the rank-order correlation between the marginal distributions of the node measures.

There are two drawbacks of this analysis which re-occur across domains. First, while this methodology tests some global relational structure of the network, we are unable to interpret the relationship between any two nodes at a high geodesic distance (≥ 2). This means that we cannot measure properties we associate with networks, such as flow or routing. Second, significance analysis is done at a particular threshold setting, without a sensitivity analysis on the original threshold choice. In the machine learning settings, the parameter sensitivity will often be on the *prediction model* parameters at a particular network definition threshold.

3.3. Neuroscience: Describing Functional Brain Structure and Their Connections

Much biological research suggests that the brain activates interconnected, often spatially distant regions along neuronal pathways [Sporns 2014]. This interconnected complexity makes networks a very natural model to study the brain. These studies are broadly in two areas: structural and functional brain networks. *structural net-*



(a) Betweenness vs. Degree (b) Betweenness vs. Closeness

Fig. 6: From [Donges et al. 2009a] (a) the distribution of betweenness centrality of nodes vs. degree of nodes. (b) the distribution of betweenness centrality of nodes vs. closeness centrality of nodes.

works (also known as 'effective connectivity', tractography, or the brain connectome) map the physical axon pathways between neurons, which may be relatively long and spatially distant. With some simplification, these networks are analogous to the physical layer in communication networks, where nodes are *explicitly* connected by cables and routers. *Functional networks* are analogous to the logical layer in communication networks. These networks model *how* neuronal signals (i.e. 'traffic') flows over this physical layer in order to activate other neurons (i.e. 'resources') to perform different types of behavior such as auditory, visual, or motor behaviors. Unfortunately, researchers do not fully understand the underlying routing and information-seeking on this physical network, nor the complex contexts which change how the behavior is realized within the functional layer. Researchers aim to better understand and predict this routing, and the collective activation dynamics in different areas of the brain.

Much work compared the topology of structural and functional networks using descriptive network statistics, and higher-order structures such as cluster and communities [Reijneveld et al. 2007; Rubinov and Sporns 2010; Sporns and Betzel 2016], especially under different experimental conditions which may affect these structures such as spinal cord injuries, epilepsy, or schizophrenia. However, all of these studies infer the network models differently, therefore it is an open challenge to rigorously synthesize these results.

3.3.1. Underlying Data in Brain Networks. The underlying data for structural or function networks are primarily derived from two sources. First, biomedical imaging technologies including Magnetic Resonance Imaging (MRI), functional MRI (fMRI), Diffusion Tensor Imaging (DTI) detect structure of varying densities and water content. These procedures produce a flat, 2-dimensional image of pixels, or a 3-dimensional volume of voxels (often as a time series of samples). For example, Diffusion Tensor Imaging (DTI) is used to construct structural networks. These images can accurately trace axon tissue connectivity by measuring flow vector orientation through the voxel space. Functional MRI (fMRI) similarly measures blood flow to voxels, a surrogate for 'activity' at this location. Inferring a functional network on fMRI data then amounts to measuring statistical interactions between activations in different brain areas. Second, non-invasive sensors including Electroenchaphalography (EEG) and Magnetoencephalography (MEG) measure and localize electrical current at a probed location.

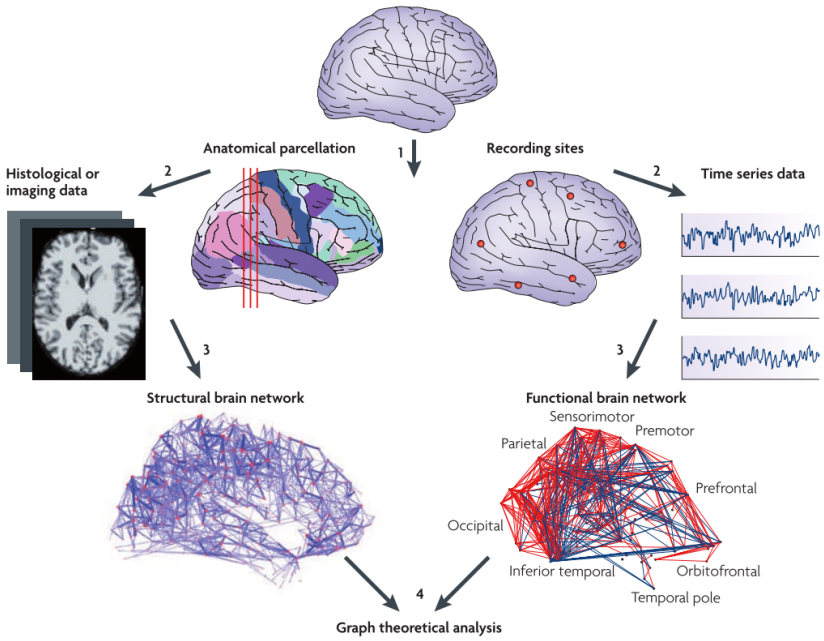


Fig. 7: From [Sporns 2014]. A data workflow for constructing brain networks. (1) for a subject or population of subjects, (2) sensing techniques such as Diffusion Tensor Imaging (DTI) or functional Magnetic Resonance Imaging (fMRI) are used to sense connectivity or activity of the brain, respectively. (3) Given sensed data at recording sites (nodes), edges are inferred by different measures on the underlying data (see: Section 3.3.2). (4) Subsequent scientific study is conducted, using the network as a data model.

Typically, these probes yield fewer and spatially coarser nodes than those defined from fMRI voxel data. These techniques produce time-series estimating electrical current (another surrogate for 'activity') at reference locations.

Figure 7 illustrates defining both structural and functional networks, from (1) the data collection on individual subjects to (4) the final analysis task. The left path of Figure 7 illustrates constructing structural networks. (2) *Anatomical parcellation* techniques use DTI or similar imaging to determine physical connectivity in the brain. These techniques are very accurate in recovering tracts of connectivity, unambiguously. (3) these tracts are translated into nodes and edges, where previous work shows significant effects of node definition on descriptive measures such as average path length and clustering coefficient [Zalesky et al. 2010]. Finally, (4) researchers use the networks as models to ask questions about the brain of the original subject or population.

The right path of Figure 7 illustrates inferring functional networks from sensed neuronal activity. This activity can be for a range of stimulus such as music preferences [Wilkins et al. 2014], image/language associations [Papalexakis et al. 2014] or for experimental conditions such as an Alzheimers patient cohort [Supekar et al. 2008a]. (2) fMRI, EEG, or MEG sensors measure activity at different *recording sites* (contact locations, pixel or voxel locations). As in structural network construction, some aggregation or node definition mapping may be applied for defining this time series dataset. (3) These activity response signals are compared between recording sites (nodes) with

time series similarity measures such as cross-correlation. 'Sufficiently' similar time series dynamics are interpreted as latent connections between nodes, yielding the final functional network.

Aside from structural and functional networks from fMRI coupled with a particular experimental user task (e.g. speaking, listening, motor), work has focused on inferring *resting*-state networks (RSNs) of the brain [Greicius et al. 2003]. These networks are constructed in much the same way as other functional networks, except this resting connectivity is informative of very robust functional clusters. Functional networks for different user tasks can then be characterized at a higher level (e.g. cognitively 'difficult' tasks) by comparing to the resting-state network (RSN).

Another network of interest is the 'rich-club' structural sub-networks [van den Heuvel et al. 2012; van den Heuvel and Sporns 2011, 2015]. This network is essentially a k-core decomposition of the structural network, which indicates the global 'backbone' of connectivity (where $k > 10$ is set in comparison to degree-preserving randomized networks). Nodes within the rich-club network are also used to characterize the broader network into 'rich-club edges' connecting two rich-club nodes, 'feeder edges' connecting exactly one rich-club node, and 'local edges' which connect two non rich-club nodes (see: Figure 8). Analogous to routing in communication networks, information can flow locally within one local region for a particular behavior, or routed through backbones to physically distant regions.

3.3.2. Methods for Inferring Networks. Time series are the dominant underlying data in neuroscience, therefore methods for constructing functional brain networks are almost exclusively in the domain of thresholded pairwise similarity measures, with some exceptions of parametric network modeling [Klimm et al. 2014]. Sakkalis [2011] provides an in-depth review of these different measures, including cross-correlation [Bialonski et al. 2011; Zalesky et al. 2012], frequency domain analysis such as discrete Fourier transform (DFT) and discrete wavelet transform (DWT) and domain-driven 'coherence' measures [Lachaux et al. 2002; Pfurtscheller and Andrew 1999; Ponten et al. 2016; Zhan et al. 2006]. Finally, significant work has focused on causal models [Ramsey et al. 2010], including Granger causality [Dhamala et al. 2008; Roebroeck et al. 2005] and dynamic causal models (DCM) [David et al. 2008; Friston et al. 2011; Rosa et al. 2012].

Where these methods have threshold parameters, τ, these methods are often validated by measuring robustness of network statistics across varying thresholds [Kramer et al. 2008], and using these thresholds for distinguishing patient cohorts by label or network statistic distribution. For example, previous work uses paired t-test or other simple statistical test [Supekar et al. 2008b]. Kramer et al. [2009] proposes a bootstrapping [Efron and Tibshirani 1993] method in the frequency domain which can provide more general p-values without model assumptions.

3.3.3. Dynamic Functional Brain Networks. Because the data underlying functional brain networks is often time series over a fixed set of nodes, a time series of networks (dynamic networks, or time-evolving networks) are a natural extension in this domain [Hutchison et al. 2013]. Network construction using time-series similarity measures (e.g. cross-correlation) generalize to the dynamic setting, computing on time-series subsequences. The advantage of introducing the complexity of dynamics is discovering distinct connectivity 'states' over the course of the experiment. Because fMRI response can change very quickly as activity occurs over the brain, these states are lost under global time series measures [Damaraju et al. 2014; Robinson et al. 2015; Yu et al. 2015]. Challenges of network validation generalize to this dynamic setting, with the added challenge of appropriate temporal *scale*.

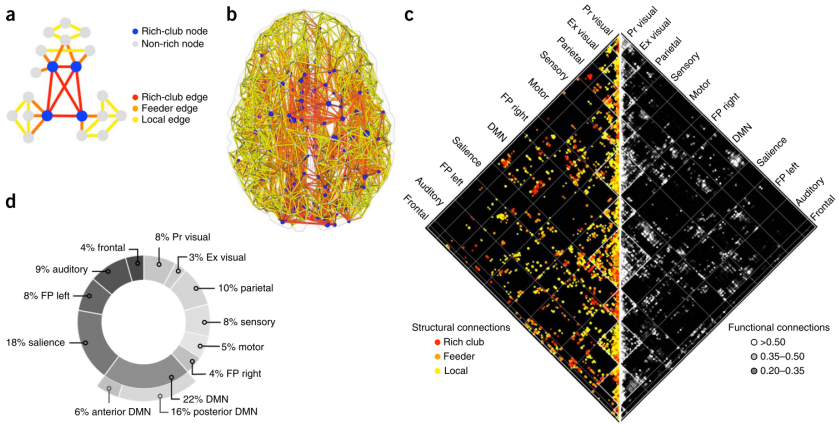


Fig. 8: From [Sporns 2014]. A study workflow comparing structural and functional brain networks. (a) a schematic of the rich-club structural network. Blue nodes indicate the k-core decomposition of the structural network ($k > 10$), grey nodes indicate non-rich nodes. Red ('rich-club') edges connect two rich-club nodes, orange ('feeder') edges connect exactly one rich-club and one non-rich node, yellow ('local') edges connect two non-rich nodes. (b) the network with colored nodes and edges, visualized in a brain coordinate system. (c) an adjacency matrix comparing structural topology sensed from Diffusion Tensor Imaging with rich-club edge coloring (left) to three thresholded values of functional connections for resting brain state inferred from fMRI for the same node-set (right). These nodes are ordered according to brain function in different spatial regions of the brain (e.g. default mode network (DMN), motor, auditory, frontal). (d) the distribution of rich-club nodes within these different labeled regions.

3.3.4. A Workflow for Comparing Functional and Structural Brain Networks. Figure 8 illustrates a complete case study which summarizes many of the topics discussed above. This work integrates structural networks across 75 individuals, sensed from Diffusion Tensor Imaging (DTI) with functional networks sensed from fMRI in resting state using Pearson correlation. These different views of the network enable researchers to study how function and connectivity are correlated. (a) illustrates a schematic layout of *rich-club* nodes, feeder, and local nodes. (b) shows the layout of these nodes in a brain coordinate system, with the same node and edge coloring. (c) illustrates an adjacency matrix comparing structural connections (left) with three thresholded values of functional connections within the same node-set (right). Furthermore, nodes in different spatial regions of the brain (e.g. default mode network (DMN), motor, auditory, frontal) are labeled according to primary function, showing structural and functional edges between these regions. (d) shows the distribution of rich-club nodes within these different labeled regions.

3.4. Epidemiology, Blogs, Information Networks: Modeling Virus Spread and Information Flow

3.4.1. Networks in Epidemiology. Networks are used in epidemiology to simulate the spread of disease over a family of parametric network models (e.g. random, small-world, exponential random graphs) representing contact between entities over time [Keeling and Eames 2005]. Network structure inference in epidemiology and information networks aims to discover an unobservable network (e.g. physical contact networks, sexual networks, malware transmission) over which information or infection is

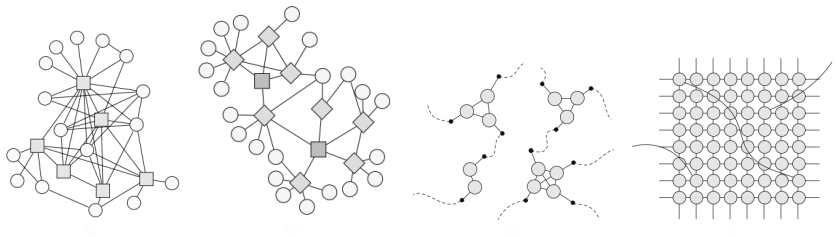


Fig. 9: Examples of networks inferred from data, or modeled in the epidemiology domain, re-printed from [Danon et al. 2011]. (a) depicts a contact network where squares are respondents, and edges are sexual or drug contact which might transmit HIV [Bell et al. 1999]. (b) a sexual network derived from snowball sampling of respondents (squares), where the edges between non-respondents (circle nodes) are unknown. (c) a network of households (cliques) and their interconnections for modeling infection in realistic social contact networks. (d) The 'small-world' network property modeled through a lattice with sparse edges connecting distant nodes.

spread. We observe the *effects* of transmission on the infected node (e.g symptoms), but not the edges over which they spread. Our focus is therefore on parametric inference for these models from non-network data.

Modeling *contact networks* allows researchers to simulate different outbreaks on these networks. Figure 9 illustrates types of network data and models used in the epidemiology domain [Danon et al. 2011]. Figure 9(a) visualizes a contact network from survey data, where squares are respondents, and edges identify transmission risk for HIV through contact by drug use or sex [Bell et al. 1999]. Figure 9(b) shows a snowball sample of respondents (as squares) and their partners. In this example, edges between circle nodes are unknown. Subsequent modeling for edges between circle nodes can test the spread over this population under different unobserved contact assumptions. Figure 9(c) illustrates a model of households (cliques), sparsely connected to others. This is intended to model contact networks and potential spread through family-unit environments. Finally, 9(d) illustrates a lattice network with sparse edges outside of the local neighborhood. This model was previously used to capture the 'small-world' property of information and disease spread [Boots and Sasaki 1999].

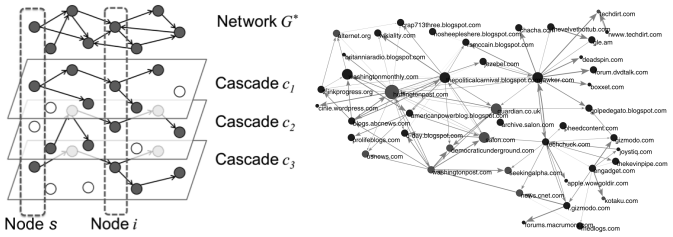
Historically, inferring networks in this area focuses largely on parametric graphical modeling using MCMC (from the epidemiology domain) and maximum likelihood methods (from machine learning), incorporating modeling assumptions in transmission rate decay. Given a transmission model (e.g. the susceptible-infected-recovered SIR model or susceptible-exposed-infected-recovered SEIR model), these methods measure the likelihood of possible *sequences*, or trees of infection, where infection times from a root are monotonically increasing. Simply, let t_i and t_j denote the infection times of nodes n_i, and n_j then the transmission model measures the probability $P(\text{“j infected i”}|t_j - t_j)$.

3.4.2. Contact Network Inference. Early work in the epidemiology domain focused the inference of either spread parameters (e.g. infection rate) or network model parameters on random graphs [Britton and O'Neill 2002], Poisson, and power-law networks [Meyers et al. 2005], and fitting of real-world data to a contact network model [Bansal et al. 2007]. In these analytical and simulation results, the network model is known and no structural inference is necessary. These studies generally model the spread

of an epidemic under possible individuals in contact (called 'contact tracing') [Patrick et al. 2002], and discovering the root individual(s) of the infection (called 'transmission tracing'). Early work also formulates association network heuristics based time and distance of potential contacts [Haydon et al. 2003]. However, this area is largely focused on parametric network modeling.

3.4.3. Infection-Time Cascades. Previous work in machine learning uses network structure inference to represent the spread of information between nodes, where the edges of transmission are unobservable. Maximum-likelihood formulations have focused on learning a network under assumed transmission rate models, using statistical inference for these parameters Myers and Leskovec [2010].

Figure 10(a) illustrates the intuition of network construction by information propagation through unobservable edges. To recover the unobservable true network G^*, each sequence of non-decreasing infection times (e.g. "cascades") supports the possible transmission between nodes with sufficiently close infection times. The key modeling step of this area of work is specifying (or learning the parameters of) a transmission model which measures the likelihood of a cascade according to differences in adjacent infection times.



(a) (b)

Fig. 10: From [Gomez-Rodriguez et al. 2012]. (a) illustrates the intuition of constructing network G^* on a collection of cascades $c_1, c_2 ... c_k$. For a possible source node s, these methods estimate the likelihood of infecting i given the observation of information arrival ('infection') at node i. (b) An estimated information network inferred from hyperlink arrival times at nodes.

Myers and Leskovec [2010] ("CONNIE") uses convex programming to learn a maximum-likelihood network under a fixed transmission time probability distributions $w(t)$, and recovery-time distribution $r(t)$. To learn the transmission weight matrix A, the authors use an Independent Cascade model [Kempe et al. 2003] where an uninfected node n_i is exposed to infection by adjacent infected nodes n_j at each time step using a Bernoulli process with a probability $A_{j,i}$. The authors present a convex optimization formulation of their likelihood function, with regularization. The CONNIE model, and most of the subsequent work, is evaluated on synthetic network models where the underlying network is explicitly known. The 'task' is the accurate reconstruction of the network which simulated these infection times. Similar to gene regulatory networks, the final evaluation of the network is the network itself, rather than a subsequent task on the inferred network. These methods then also typically present qualitative results on real-world datasets.

CONNIE uses a geometric program and is not scalable. Gomez-Rodriguez et al. [2012] ("NETINF") solves a simplified problem in a scalable way by fixing a global edge

transmission probability β. For many applications, this assumption of a fixed transmission threshold can be made. The primary insight under this assumption is that we can simply use the most likely propagation tree over a set of nodes in a cascade c. Given a cascade set C, the authors marginalize their likelihood function relative to edge selection and prove this function is monotonic and submodular. Therefore, edges can be greedily selected with an approximation factor of $(1 - 1/e)$ [Nemhauser et al. 1978].

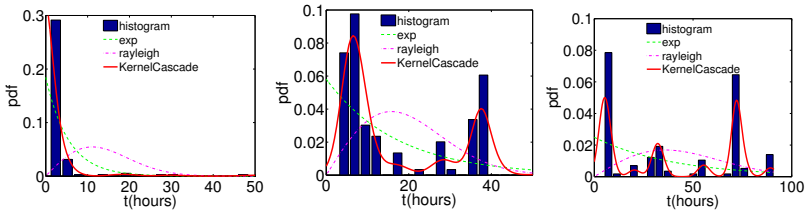


Fig. 11: From [Du et al. 2012]. Three observed transmission delay histograms, demonstrating poor fit of exponential and Raleigh models.

The transmission model of [Myers and Leskovec 2010] assumes repeated Bernoulli trials of fixed probability. Gomez-Rodriguez et al. [2014] ("NETRATE") introduces transmission likelihood functions which vary over time. Given a set of cascades C, the method infers the parameters of transmission rate models for each edge. The model uses a hazard function $H(\tau_i|\tau_j;\alpha_{j,i})$ which measures the instantaneous infection rate on node n_i from infected node n_j, given the parametric function on α. The authors present three different transmission model functions: Exponential, Power-law, and Rayleigh, and prove that the optimization for transmission rates matrix A is convex.

In many real online applications, information can propagate in a multitude of ways. Information might be promoted by an influential node, causing multimodal spikes in propagation after some delays. These transmission rates may not decay monotonically, nor according to any parametric function. Figure 11 illustrates this property on real transmission data between blog sites in the MemeTracker[4] dataset. The x-axis is the time difference between post creation on the originating site, and the time it was linked by another site. The data poorly fits any single exponential or Rayleigh transmission model. Du et al. [2012] ("KernelCascade") extends NETRATE to address this limitation. The key addition of this work is to kernelize the hazard function $H(\tau_*|\tau_j,\alpha_{j,i})$ over m different kernels. These kernels serve as a piecewise approximation of the time-lag distribution, which can then be used to estimate the likelihood of transmission between nodes, given observed 'infection' data.

3.5. Ecology: Inferring Animal Social Networks to Explain Individual and Group Behavior

3.5.1. Networks in Ecology. Networks in ecology serve as two distinct models. In systems ecology, traditional graphical models are used to model an ecosystem at a high level, with relationships between species, environmental variables, services, and other processes [Milns et al. 2010]. For example, food webs model who-eats-whom within an ecosystem [Proulx et al. 2005]. A second modeling with networks–and the focus of this section–arises in behavioral ecology for the study of animal populations [Farine and Whitehead 2015; Wey et al. 2008]. Analogous to sociology and political science, traditional fieldwork data in ecology are collected from direct observation and 'surveys,'

[4]http://www.memetracker.org/

measuring interactions or other relationships among individuals in the population [Barrett et al. 2012; Sueur et al. 2011], particularly over time [de Silva et al. 2011; McDonald 2007; Pinter-Wollman et al. 2013].

This process of direct observation allows researchers to incorporate their own intuition and experience into the definition of these networks. In practice, much of the work in this area uses ad-hoc, intuitive network definitions with some sensitivity analysis. These networks derived from direct observation are typically categorical (e.g. kinship relations) or discrete (e.g. thresholding on number of interactions, where interactions are implicitly decided by observers). Methodological consideration are well established in the field on these data, including edge strength thresholding [Croft et al. 2009], sampling, hypothesis testing [Croft et al. 2011], and randomization strategies [Haddadi et al. 2011; James et al. 2009]. Each of these provide several different choices for validating the robustness of these networks. These networks are tightly coupled with a particular hypothesis and experimental cohorts in the population.

Two networks of interest in animal social networks measure *affiliations* and *associations* between individuals [Whitehead and James 2015]. Affiliations describe *intentional* social relationships between individuals [Croft et al. 2011; Whitehead et al. 2005] (e.g. grooming pairs of primates), while associations describe a broader set of interactions which might be driven by structural factors rather than social affinity (e.g. environmental resources, sex, age, and other individual attributes) [Bejder et al. 1998]. Whitehead and James [2015] introduces a generalized affiliation index using a linear regression model. The authors simplify the model by subtracting sets of predictive structural features to be removed. The resulting model by subtraction is the affiliative network model.

3.5.2. Underlying Data in Animal Social Networks: Instrumentation and Sensing of Animal Populations. Recent instrumentation of individuals and the environment allow the observation of ecosystems and populations at an unprecedented scale using geo-location sensors such as GPS, proximity sensors, radio-telemetry, and Passive Integrated Transponder (PIT) tags [Kays et al. 2015; Krause et al. 2013; Rutz et al. 2012]. This instrumentation allows the study of detailed individual behavior and social dynamics which are often outside of the view of direct observation. This abundance of data requires novel statistical techniques for inferring networks from *implicit* interactions. Because these studies are often coupled with an experimental design, these individual labels (e.g. test and control populations) are often used in visualization and hypothesis testing.

No known work compares the biases of interaction and/or affiliation sampling via traditional fieldwork, against the capability of simultaneous sensing for collecting underlying data in animal social networks. Presently, these sensors are most effective at recording simple co-location or trajectories. Challenging independent problems such as activity recognition (e.g. grooming, conflict) are more easily solved by researchers doing direct observation on the population. Yet, researchers are naturally limited in their attention and accuracy. Future research will likely integrate the strengths of these modalities to augment fieldwork data collection.

3.5.3. Studies and Network Inference Methods on Instrumented Data. The key difference between data from traditional fieldwork and from instrumented technologies is that the former tend to be discrete counts (e.g. number of co-locations or grooming events), while the latter are continuous data without these higher-level labels (e.g. relative distances between individuals). To translate to discrete co-location events–and subsequently a network–requires defining "how close" for "which duration" constitutes a co-location edge, or "how correlated" for "which duration" constitutes a "following" edge in the network. In contrast, researchers easily identify these relationships informally.

The simplest method for setting these closeness and persistence thresholds for co-location (i.e. measuring 'association' networks as described above) is by domain knowledge, or by sampling the parameter-space in some way. Haddadi et al. [2011] use this strategy in GPS data from sheep, ranging from individuals co-locating for 1 minute at 1 meter, to 5 minutes at 3.5 meters. The authors have some known 'affiliations' (as described above), which are used to validate network accuracy at these different thresholds when the individuals are mixed into a larger population. Aplin et al. [2012] collect data from passive integrated transponder (PIT) tags of individuals sensed by radio-frequency identification (RFID) antennae at feeder sites. This work defines associations as two individuals co-occurring at the site within 30 seconds before or after the other on a sliding 75 second window. Co-occurrence is categorical due to the physical design of feeders, so only the 'persistence' of interaction need be fixed. This threshold generates a stream of pairwise associations which can then be thresholded ($\tau \geq 0.02$) to produce an aggregated association network.

Psorakis et al. [2012] define edges using Gaussian mixture models (GMM) on co-occurrence data for a similar feeder system. This approach mitigates the 'persistence' threshold by fitting Gaussian distributions to a one-dimensional space of occurrence counts (and later continuous two-dimensional geographic space, [Farine et al. 2016]). These distributions capture *events* of co-occurrence among several individuals. Hamede et al. [2009] use a randomization approach to define non-random associations on proximity sensors on a population of wild Tasmanian devils (*Sarcophilus harrisii*). Internal thresholds on these sensors detect co-location within 30 centimeters of each other. This work studies disease transmission through physical contact of the animals, so this thresholding is appropriate.

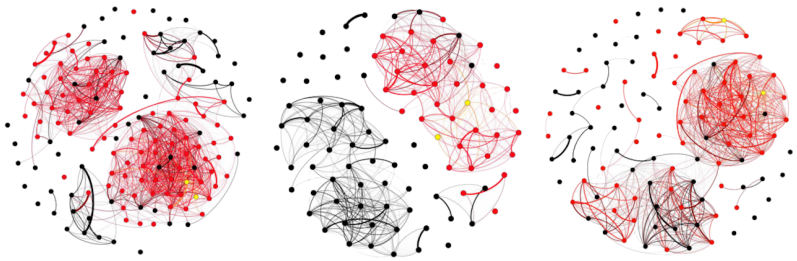


Fig. 12: From [Aplin et al. 2015], Networks inferred from co-location at feeding stations sensed using RFID over three different populations. These edges are colored showing learned behavior (red edges) of obtaining food through an instrumented puzzle mechanism, the trained individuals (yellow nodes), and affiliated individuals using the default strategy (black edges).

3.5.4. A Workflow from Co-Location Data to Innovation Spread in Networks. Aplin et al. [2015] proposes a network inference task to measure the learning of a feeding behavior in great tits (*Parus major*). In this experiment, feeders instrumented with RFID antennae record the visitation of each unique bird using PIT tags. The feeders use a sliding door to the left or right to access the food, and this feeder records the bird's puzzle solution. The authors investigate whether birds learn by example at the feeding sites. Figure 12 shows a thresholded, aggregated network over individuals, weighted by the frequency of co-location events at any feeder, using the Gaussian mixture model

(GMM) method described above for interaction 'events.' Yellow nodes represent trained individuals, red nodes represent individuals learning the correct behavior to solve the feeder by the end of the study.

These networks visually show strong network modules between red and black individuals. Figure 12(b) shows a strong network separation between the two behaviors, where trained individuals are within the red cluster. Figure 12(c) shows two strong red clusters around both trained individuals, but also that the correct behavior is spread across a component of untrained individuals.

3.6. Mobile Social Networks: Studying Human Mobility Through Social Relationships

3.6.1. Underlying Data in Mobile Networks. Phones and other mobile devices are among the versatile and informative sensors of personal and social activity [Lane et al. 2010]. The modeling of mobile phone data as networks is motivated by the complex, overlapping, and dynamic modalities which are sensed by these devices. Mobile devices collect physical proximity (bluetooth, WiFi), physical location (GPS), direct communication (SMS, voice), and often interactions through other online social networks (e.g. Facebook).[5] Integrating these modalities promises to give a rich picture of large scale human mobility, dynamics and scale [Saramäki and Moro 2015], geography and communication [Blondel et al. 2010; Ratti et al. 2010], and offline social networks.

While many of these underlying data are very similar to those collected for animal social networks (proximity, location, discrete interactions), there are notable trade-offs between privacy and experimental design in these domains. While animals are not due rights to data privacy, they are also unable to comply with instructions or be surveyed for ground truth network edges; in human experiments, contact diaries [Mastrandrea et al. 2015] or Facebook friends [Sekara and Lehmann 2014] have been collected to validate networks inferred from proximity sensors. Experiments on mobile users are necessarily less invasive, while topics such as disease spread and sexual contact networks are often sensed in animal populations. Data privacy requires careful, informed consent and secure storage [Stopczynski et al. 2014a]; location privacy has been shown as extremely identifiable, with two to eleven data points being sufficient to uniquely identify individuals [de Montjoye et al. 2013].

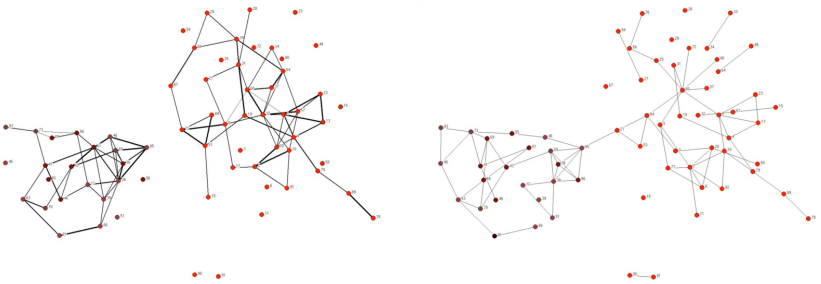
Several mobile datasets have been collected for the purposes of social research [Blondel et al. 2015]. The first large collection was the "Reality Mining" dataset, collected on 100 participants (faculty and students) in the MIT Media Laboratory [Eagle and (Sandy) Pentland 2006]. This anonymized dataset contains call logs, bluetooth device proximity, cell tower ID (a proxy for location) and other fields. Similar mobile data collection projects followed, including the Lausanne Data Collection Campaign on 170 student participants [Laurila et al. 2013], the Social fMRI study on 130 participants [Aharony et al. 2011] (notably, not university students), and the SensibleDTU project of $1,000$ participants [Stopczynski et al. 2014b]. These subsequent studies collected more detailed user activity, surveys, Facebook, and detailed user demographics, addressing the limitations of previous efforts.

Finally, the SocioPatterns platform [Barrat et al. 2008; Cattuto et al. 2010] uses a specialized proximity sensor design to record face-to-face interactions. These sensors have been deployed in an academic conference setting [Smieszek et al. 2016], elementary schools [Stehl et al. 2013], high schools [Mastrandrea et al. 2015] and several other environments.[6] The specificity of these sensors for detecting individual *interac-*

[5] In this section, we also group email datasets such as Enron [Klimt and Yang Klimt and Yang] because the discrete interaction data is most similar to this domain.
[6] http://www.sociopatterns.org/datasets/

tions between users addresses the challenges of using general proximity sensors (e.g. bluetooth) for population studies.

3.6.2. Studies and Methods on Mobile Data. The most common primary task in inferring networks from mobile data is related to *integration* across modalities, for edge or attribute prediction tasks. For example, predicting Facebook friends from bluetooth co-location [Sekara and Lehmann 2014], or predicting survey-reported friends from proximity and call record data [Eagle et al. 2009].



(a) Network inferred from user co-location and call data.

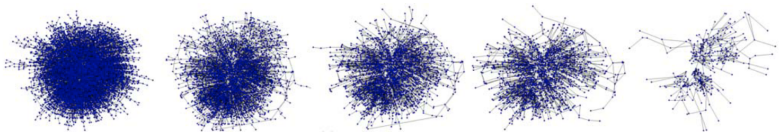
(b) The social network self-reported from a user survey.

Fig. 13: From [Eagle et al. 2009]. (a) a weighted latent network inferred by bluetooth proximity and call records on the population of Reality Mining users. (b) the self-reported ground-truth friendship network.

Figure 13 examines this latter task. The network in Figure 13(a) is inferred using principle components analysis of bluetooth proximity counts across different times and locations (e.g. work, off-campus, weekday, weekend), and assigning edge weight by the coefficient of the factor corresponding to *non-work* hours (e.g. "close friends are those co-located outside of work"). Figure 13(b) reports the ground-truth social network, self-reported from a user survey, accurately reconstructed by the inferred network. While the discovery that friends meet or call after work is not particularly surprising, this demonstrates integrating these modalities for the simple edge prediction task. The principle components measure to infer the network edges also incorporates domain knowledge of work schedules. Previous work shows that human mobility in urban environments is highly periodic between a small set of locations (e.g. home and work) [Eagle et al. 2009]. Therefore, incorporating these periodicities explicitly is a key aspect of this domain.

De Choudhury et al. [2010] revisits the discussion of setting similarity threshold τ for an interaction measure calculated on data. Whether the underlying data is of discrete or continuous, varying τ realizes a range of possible networks. In Figure 14, from left to right the number of required emails exchanged increases in order to define an edge, thus reducing the density. Any predictive task on this network balances novelty against the task difficulty: at a low threshold, a dense graph is realized and edge prediction may not perform better than random because the definition of the edge is simply noise. However, a very high threshold may infer a very sparse network, where edges are trivially easy to predict (but uninteresting).

The authors generate two and four one-year aggregated networks according to the length of data available in a university email dataset, and the Enron email dataset.



(a) A network inferred from a university email corpus

(b) A network inferred from the Enron email corpus

Fig. 14: From [De Choudhury et al. 2010]. Two networks inferred over varying network thresholds τ, creating sparser networks from left to right as the threshold criteria becomes more strict (e.g. more than τ email interactions for an edge between individuals v_i and v_j).

They then tune the global threshold τ according to performance across several different classification tasks on sets of node-level features from each inferred network. These tasks include classification of node class (e.g. undergraduate, graduate, faculty, staff), gender, "community" (with class labels provided by stochastic block modeling, [Hofman and Wiggins 2008]), where each of these tasks may be independently of interest. Assuming these node demographics and communities were separable as a set of behaviors at some "natural" threshold, this analysis will discover the threshold yielding the maximal classification accuracy. The authors also predict future communication activity using simple linear regression, reporting the accuracy at these same τ. Each of these tasks yield a similar range of high performing τ thresholds, suggesting that classification and prediction agree across multiple views of the network.

4. CONCLUSION

This survey aims to provide a vocabulary and structure to the problem of inferring networks from non-network data. Typically, this problem is addressed in data preprocessing, often with several *artful* steps of parameter tuning or feature selection. We anticipate investigation of this problem in a more general and rigorous framework as network models from underlying non-network data are more numerous in data science applications.

We survey several domains in order to illustrate the different foci in questions and how the nature of data drive the methodological specializations in the areas. For example–with some simplification–we observe that gene regulatory networks are methodologically very mature, with a breadth of interaction measures appropriate for multivariate, *matrix* data (e.g. microarray) including regression and graphical models. Climate networks and brain networks are mature in *time series* interaction measures, including causal and frequency-based analysis, respectively. The problems in each of these areas are still exploratory, focusing on integrating and validating networks from different data (e.g. structural and functional brain networks) to develop data science tools downstream from these robust network models. Animal social networks inherit *direct observation* data in relatively simple formats (e.g. counts). This yields relatively simple network models over straightforward parameters (e.g. closeness and persistence), with a focus on experimental design. Epidemiology historically studies observed infection data *spreading* across a hidden contact network. Therefore, modeling these

transmission functions is a key to this area. We hope that this *data-driven* summary might help locate models and expertise on networks derived from different underlying data modalities.

Previous work often assumes that the objective of network inference is uncovering "the network" representation which is obscured by noise. Often in this context, the network inference method tries to reconstruct known ground-truth networks from non-network data. In contrast, our work treats a network as a model to perform a particular *task*, where we often cannot access the ground truth network, or assume its parametric form. Analogous to clustering for a classification *task*, there are many possible clusterings which are only as valuable as they improve classification accuracy. Conceptualizing network inference within the complete data science workflow–from data (to network) to task models for particular questions–focuses on a tighter coupling of data models and task models.

There are several frameworks across domains which use randomization, causality, and significance testing strategies to rigorously learn the network model under some assumptions. While these networks are appropriate according to their structural assumptions, they may not be the most informative for the question/task(s) of interest. Currently, no general, statistically rigorous, joint inference/prediction framework exists which learns a maximally predictive network model across particular task(s). Furthermore, there is little understanding of the criteria for network models and predictive models which would make them appropriate for this joint modeling. We anticipate this will be an exciting area of future research.

5. ACKNOWLEDGEMENTS

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-JRNL-703477).

REFERENCES

E Adar and L A Adamic. 2005. Tracking information epidemics in blogspace. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*. 207–214. DOI:http://dx.doi.org/10.1109/WI.2005.151

Charu Aggarwal and Karthik Subbian. 2014. Evolutionary Network Analysis: A Survey. *ACM Comput. Surv.* 47, 1 (May 2014), 10:1—-10:36. DOI:http://dx.doi.org/10.1145/2601412

Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. 2011. Social fMRI: Investigating and Shaping Social Mechanisms in the Real World. *Pervasive Mob. Comput.* 7, 6 (Dec. 2011), 643–659. DOI:http://dx.doi.org/10.1016/j.pmcj.2011.09.004

Genevera I Allen and Zhandong Liu. 2012. A Log-Linear Graphical Model for Inferring Genetic Networks from High-throughput Sequencing Data. In *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (BIBM '12)*. IEEE Computer Society, 1–6. DOI:http://dx.doi.org/10.1109/BIBM.2012.6392619

Lucy M Aplin, Damien R Farine, Julie Morand-Ferron, Andrew Cockburn, Alex Thornton, and Ben C Sheldon. 2015. Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature* 518, 7540 (Feb. 2015), 538–541. DOI:http://dx.doi.org/10.1038/nature13998

L M Aplin, D R Farine, J Morand-Ferron, and B C Sheldon. 2012. Social networks predict patch discovery in a wild population of songbirds. *Proceedings of the Royal Society of London B: Biological Sciences* (2012). DOI:http://dx.doi.org/10.1098/rspb.2012.1591

Mahmoudreza Babaei, Przemyslaw Grabowicz, Isabel Valera, Krishna P Gummadi, and Manuel Gomez-Rodriguez. 2016. On the Efficiency of the Information Networks in Social Media. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. ACM, 83–92. DOI:http://dx.doi.org/10.1145/2835776.2835826

Lars Backstrom and Jon Kleinberg. 2011. Network Bucket Testing. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. ACM, 615–624. DOI:http://dx.doi.org/10.1145/1963405.1963492

Shweta Bansal, Bryan T Grenfell, and Lauren Ancel Meyers. 2007. When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of The Royal Society Interface* 4, 16 (2007), 879–891. DOI:http://dx.doi.org/10.1098/rsif.2007.1100

Ziv Bar-Joseph, Anthony Gitter, and Itamar Simon. 2012. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics* 13, 8 (Aug. 2012), 552–564. DOI:http://dx.doi.org/10.1038/nrg3244

Alain Barrat, Ciro Cattuto, Vittoria Colizza, Jean-François Pinton, Wouter Van den Broeck, and Alessandro Vespignani. 2008. High Resolution Dynamical Mapping of Social Interactions With Active RFID. *ArXiv e-prints* (Nov. 2008). http://arxiv.org/abs/0811.4170v2

Louise Barrett, S Peter Henzi, and David Lusseau. 2012. Taking sociality seriously: the structure of multi-dimensional social networks as a source of information for individuals. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367, 1599 (2012), 2108–2118. DOI:http://dx.doi.org/10.1098/rstb.2012.0113

Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. 2013. NCBI GEO: archive for functional genomics data setsupdate. *Nucleic Acids Research* 41, D1 (2013), D991–D995. DOI:http://dx.doi.org/10.1093/nar/gks1193

Baruch Barzel and Albert-Laszlo Barabasi. 2013. Network link prediction by global silencing of indirect correlations. *Nature Biotechnology* 31, 8 (aug 2013), 720–725. DOI:http://dx.doi.org/10.1038/nbt.2601

Lars Bejder, David Fletcher, and Stefan Bräger. 1998. A method for testing association patterns of social animals. *Animal Behaviour* 56, 3 (1998), 719–725. DOI:http://dx.doi.org/10.1006/anbe.1998.0802

David C Bell, John S Atkinson, and Jerry W Carlson. 1999. Centrality measures for disease transmission networks. *Social Networks* 21, 1 (1999), 1–21. DOI:http://dx.doi.org/10.1016/S0378-8733(98)00010-0

Stephan Bialonski, Martin Wendler, and Klaus Lehnertz. 2011. Unraveling Spurious Properties of Interaction Networks with Tailored Random Networks. *PLoS ONE* 6, 8 (Aug. 2011), e22826. DOI:http://dx.doi.org/10.1371%2Fjournal.pone.0022826

Concha Bielza and Pedro Larrañaga. 2014. Bayesian networks in neuroscience: a survey. *Frontiers in Computational Neuroscience* 8 (Oct. 2014), 131. DOI:http://dx.doi.org/10.3389/fncom.2014.00131

Vincent D Blondel, Adeline Decuyper, and Gautier Krings. 2015. A survey of results on mobile phone datasets analysis. *EPJ Data Science* 4, 1 (2015), 1–55. DOI:http://dx.doi.org/10.1140/epjds/s13688-015-0046-0

Vincent D. Blondel, Gautier Krings, and Isabelle Thomas. 2010. Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies* (2010). DOI:http://dx.doi.org/2078.1/95261

S Boccaletti, G Bianconi, R Criado, C I del Genio, J Gómez-Gardeñes, M Romance, I Sendiña-Nadal, Z Wang, and M Zanin. 2014. The structure and

dynamics of multilayer networks. *Physics Reports* 544, 1 (2014), 1–122. DOI:http://dx.doi.org/10.1016/j.physrep.2014.07.001

M Boots and A Sasaki. 1999. "Small worlds" and the evolution of virulence: infection occurs locally and at a distance. *Proceedings of the Royal Society of London B: Biological Sciences* 266, 1432 (1999), 1933–1938. DOI:http://dx.doi.org/10.1098/rspb.1999.0869

Tom Britton and Philip D. O'Neill. 2002. Bayesian Inference for Stochastic Epidemics in Populations with Random Social Structure. *Scandinavian Journal of Statistics* 29, 3 (Sept. 2002), 375–390. DOI:http://dx.doi.org/10.1111/1467-9469.00296

Ed Bullmore and Olaf Sporns. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci* 10, 3 (Mar. 2009), 186–198. http://dx.doi.org/10.1038/nrn2575

Atul J Butte and Isaac S Kohane. 2000. Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements. *Pacific Symposium on Biocomputing* 5 (2000), 415–426. DOI:http://dx.doi.org/10.1142/9789814447331_0040

Leila Cammoun, Xavier Gigandet, Djalel Meskaldji, Jean Philippe Thiran, Olaf Sporns, Kim Q Do, Philippe Maeder, Reto Meuli, and Patric Hagmann. 2012. Mapping the human connectome at multiple scales with diffusion spectrum MRI. *Journal of Neuroscience Methods* 203, 2 (2012), 386–397. DOI:http://dx.doi.org/10.1016/j.jneumeth.2011.09.031

Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-Franois Pinton, and Alessandro Vespignani. 2010. Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. *PLOS ONE* 5, 7 (July 2010), e11596. DOI:http://dx.doi.org/10.1371/journal.pone.0011596

Hau Chan, Hanghang Tong, and Leman Akoglu. Make It or Break It: Manipulating Robustness in Large Networks. In *Proceedings of the 2014 SIAM International Conference on Data Mining*. 325–333. DOI:http://dx.doi.org/10.1137/1.9781611973440.37

Darren P Croft, Jens Krause, Safi K Darden, Indar W Ramnarine, Jolyon J Faria, and Richard James. 2009. Behavioural Trait Assortment in a Social Network: Patterns and Implications. *Behavioral Ecology and Sociobiology* 63, 10 (2009), 1495–1503. DOI:http://dx.doi.org/10.1007/s00265-009-0802-x

Darren P Croft, Joah R Madden, Daniel W Franks, and Richard James. 2011. Hypothesis testing in animal social networks. *Trends in Ecology & Evolution* 26, 10 (2011), 502–507. DOI:http://dx.doi.org/10.1016/j.tree.2011.05.012

Luciano da Fontoura Costa, Osvaldo N Oliveira Jr., Gonzalo Travieso, Francisco Aparecido Rodrigues, Paulino Ribeiro Villas Boas, Lucas Antiqueira, Matheus Palhares Viana, and Luis Enrique Correa Rocha. 2011. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics* 60, 3 (2011), 329–412. DOI:http://dx.doi.org/10.1080/00018732.2011.572452

E Damaraju, E A Allen, A Belger, J M Ford, S McEwen, D H Mathalon, B A Mueller, G D Pearlson, S G Potkin, A Preda, J A Turner, J G Vaidya, T G van Erp, and V D Calhoun. 2014. Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage: Clinical* 5 (2014), 298–308. DOI:http://dx.doi.org/10.1016/j.nicl.2014.07.003

Leon Danon, Ashley P Ford, Thomas House, Chris P Jewell, Matt J Keeling, Gareth O Roberts, Joshua V Ross, and Matthew C Vernon. 2011. Networks and the Epidemiology of Infectious Disease. In *Interdisciplinary Perspectives on Infectious Diseases*, Vol. 2011. DOI:http://dx.doi.org/10.1155/2011/284909

Olivier David, Isabelle Guillemain, Sandrine Saillet, Sebastien Reyt, Colin Deransart, Christoph Segebarth, and Antoine Depaulis. 2008. Identifying Neural Drivers with

Functional MRI: An Electrophysiological Validation. *PLoS Biology* 6, 12 (Dec. 2008), e315. DOI:http://dx.doi.org/10.1371/journal.pbio.0060315

Munmun De Choudhury, Winter A Mason, Jake M Hofman, and Duncan J Watts. 2010. Inferring Relevant Social Networks from Interpersonal Communication. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, 301–310. DOI:http://dx.doi.org/10.1145/1772690.1772722

Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Nature Scientific Reports* 3 (Mar 2013), 1376. DOI:http://dx.doi.org/10.1038/srep01376

Shermin de Silva, Ashoka D G Ranjeewa, and Sergey Kryazhimskiy. 2011. The dynamics of social networks among female Asian elephants. *BMC Ecology* 11, 1 (2011), 1–16. DOI:http://dx.doi.org/10.1186/1472-6785-11-17

Mukeshwar Dhamala, Govindan Rangarajan, and Mingzhou Ding. 2008. Analyzing information flow in brain networks with nonparametric Granger causality. *NeuroImage* 41, 2 (2008), 354–362. DOI:http://dx.doi.org/10.1016/j.neuroimage.2008.02.020

J F Donges, Y Zou, N Marwan, and J Kurths. 2009a. Complex networks in climate dynamics. *The European Physical Journal Special Topics* 174, 1 (2009), 157–179. DOI:http://dx.doi.org/10.1140/epjst/e2009-01098-2

J F Donges, Y Zou, N Marwan, and J Kurths. 2009b. The backbone of the climate network. *EPL (Europhysics Letters)* 87, 4 (2009), 48007. DOI:http://dx.doi.org/10.1209/0295-5075/87/48007

Nan Du, Le Song, Ming Yuan, and Alex J Smola. 2012. Learning Networks of Heterogeneous Influence. In *Advances in Neural Information Processing Systems 25*, F Pereira, C J C Burges, L Bottou, and K Q Weinberger (Eds.). 2780–2788. http://papers.nips.cc/paper/4582-learning-networks-of-heterogeneous-influence

Nathan Eagle, A.S. Pentland, and David Lazer. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* 106, 36 (Sept. 2009), 15274–15278. DOI:http://dx.doi.org/10.1073/pnas.0900282106

Nathan Eagle and Alex (Sandy) Pentland. 2006. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing* 10, 4 (Mar. 2006), 255–268. DOI:http://dx.doi.org/10.1007/s00779-005-0046-3

Imme Ebert-Uphoff and Yi Deng. 2012. Causal Discovery for Climate Research Using Graphical Models. *Journal of Climate* 25, 17 (Feb. 2012), 5648–5665. DOI:http://dx.doi.org/10.1175/JCLI-D-11-00387.1

B Efron and R J Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall.

Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. 2007. Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biology* 5, 1 (Jan. 2007), e8. DOI:http://dx.doi.org/10.1371/journal.pbio.0050008

Damien R. Farine, Ariana Strandburg-Peshkin, Tanya Berger-Wolf, Brian Ziebart, Ivan Brugere, Jia Li, and Margaret C. Crofoot. 2016. Both Nearest Neighbours and Long-term Affiliates Predict Individual Locations During Collective Movement in Wild Baboons. *Scientific Reports* 6 (June 2016), 27704. DOI:http://dx.doi.org/10.1038/srep27704

Damien R Farine and Hal Whitehead. 2015. Constructing, conducting and interpreting animal social network analysis. *Journal of Animal Ecology* 84, 5 (2015), 1144–1163. DOI:http://dx.doi.org/10.1111/1365-2656.12418

Soheil Feizi, Daniel Marbach, Muriel Medard, and Manolis Kellis. 2013. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature Biotechnology* 31, 8 (Aug. 2013), 726–733.

DOI:http://dx.doi.org/10.1038/nbt.2635
Karl J Friston, Baojuan Li, Jean Daunizeau, and Klaas E Stephan. 2011. Network discovery with DCM. *NeuroImage* 56, 3 (Jun. 2011), 1202–1221. DOI:http://dx.doi.org/10.1016/j.neuroimage.2010.12.039
Lise Getoor, Nir Friedman, Daphne Koller, Avi Pfeffer, and Ben Taskar. 2007. Probabilistic Relational Models. *Introduction to Statistical Relational Learning* (2007), 129. https://mitpress.mit.edu/books/introduction-statistical-relational-learning
Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. 2010. A Survey of Statistical Network Models. *Foundations and Trends in Machine Learning* 2, 2 (Feb. 2010), 129–233. DOI:http://dx.doi.org/10.1561/2200000005
Manuel Gomez-Rodriguez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf. 2014. Uncovering the structure and temporal dynamics of information propagation. *Network Science* 2, 1 (4 2014), 26–65. DOI:http://dx.doi.org/10.1017/nws.2014.3
Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. 2012. Inferring Networks of Diffusion and Influence. *ACM Transactions on Knowledge Discovery from Data* 5, 4 (Feb. 2012), 21:1—-21:37. DOI:http://dx.doi.org/10.1145/2086737.2086741
Neil Zhenqiang Gong, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine (Runting) Shi, and Dawn Song. 2014. Joint Link Prediction and Attribute Inference Using a Social-Attribute Network. *ACM Trans. Intell. Syst. Technol.* 5, 2 (Apr. 2014), 27:1—-27:20. DOI:http://dx.doi.org/10.1145/2594455
C W J Granger. 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* 37, 3 (1969), 424–438. DOI:http://dx.doi.org/10.2307/1912791
Michael D Greicius, Ben Krasnow, Allan L Reiss, and Vinod Menon. 2003. Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences* 100, 1 (2003), 253–258. DOI:http://dx.doi.org/10.1073/pnas.0135058100
Chris Groendyke, David Welch, and David R. Hunter. 2011. Bayesian Inference for Contact Networks Given Epidemic Data. *Scandinavian Journal of Statistics* 38, 3 (Sept. 2011), 600–616. DOI:http://dx.doi.org/10.1111/j.1467-9469.2010.00721.x
Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. 2015. Network A/B Testing: From Sampling to Estimation. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. ACM, 399–409. DOI:http://dx.doi.org/10.1145/2736277.2741081
Hamed Haddadi, Andrew J. King, Alison P. Wills, Damien Fay, John Lowe, A. Jennifer Morton, Stephen Hailes, and Alan M. Wilson. 2011. Determining association networks in social animals: choosing spatialtemporal criteria and sampling rates. *Behavioral Ecology and Sociobiology* 65, 8 (2011), 1659–1668. DOI:http://dx.doi.org/10.1007/s00265-011-1193-3
H Haddadi, M Rio, Gianluca Iannaccone, A Moore, and R Mortier. 2008. Network topologies: inference, modeling, and generation. *Communications Surveys Tutorials, IEEE* 10, 2 (2008), 48–69. DOI:http://dx.doi.org/10.1109/COMST.2008.4564479
Rodrigo K Hamede, Jim Bashford, Hamish McCallum, and Menna Jones. 2009. Contact networks in a wild Tasmanian devil (Sarcophilus harrisii) population: using social network analysis to reveal seasonal variability in social behaviour and its implications for transmission of devil facial tumour disease. *Ecology Letters* 12, 11 (2009), 1147–1157. DOI:http://dx.doi.org/10.1111/j.1461-0248.2009.01370.x
Mohammad Al Hasan and Mohammed J. Zaki. 2011. A Survey of Link Prediction in Social Networks. In *Social Network Data Analytics SE - 9*, Charu C Aggarwal (Ed.). Springer US, 243–275. DOI:http://dx.doi.org/10.1007/978-1-4419-8462-3_9
Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-

Philippe Vert. 2012. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Systems Biology* 6, 1 (2012), 145. DOI:http://dx.doi.org/10.1186/1752-0509-6-145

S Havlin, D Y Kenett, E Ben-Jacob, A Bunde, R Cohen, H Hermann, J W Kantelhardt, J Kertész, S Kirkpatrick, J Kurths, J Portugali, and S Solomon. 2012. Challenges in network science: Applications to infrastructures, climate, social systems and economics. *The European Physical Journal Special Topics* 214, 1 (2012), 273–293. DOI:http://dx.doi.org/10.1140/epjst/e2012-01695-x

D. T. Haydon, M. Chase-Topping, D. J. Shaw, L. Matthews, J. K. Friar, J. Wilesmith, and M. E. J. Woolhouse. 2003. The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proceedings of the Royal Society of London B: Biological Sciences* 270, 1511 (Jan. 2003), 121–127. DOI:http://dx.doi.org/10.1098/rspb.2002.2191

Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene van Someren, and Reinhard Guthke. 2009. Gene regulatory network inference: Data integration in dynamic models–A review. *Biosystems* 96, 1 (2009), 86–103. DOI:http://dx.doi.org/10.1016/j.biosystems.2008.12.004

Anthony Hey, Stewart Tansley, and Kristin Tolle. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

Jaroslav Hlinka, David Hartman, Martin Vejmelka, Jakob Runge, Norbert Marwan, Jürgen Kurths, and Milan Paluš. 2013. Reliability of Inference of Directed Climate Networks Using Conditional Mutual Information. *Entropy* 15, 6 (2013), 2023. DOI:http://dx.doi.org/10.3390/e15062023

Jake M Hofman and Chris H Wiggins. 2008. Bayesian Approach to Network Modularity. *Physical Review Letters* 100, 25 (Jun. 2008), 258701. DOI:http://dx.doi.org/10.1103/PhysRevLett.100.258701

Petter Holme. 2015. Modern temporal network theory: a colloquium. *The European Physical Journal B* 88, 9 (2015), 1–30. DOI:http://dx.doi.org/10.1140/epjb/e2015-60657-4

Petter Holme and Jari Saramäki. 2012. Temporal networks. *Physics Reports* 519, 3 (2012), 97–125. DOI:http://dx.doi.org/10.1016/j.physrep.2012.03.001

R Matthew Hutchison, Thilo Womelsdorf, Elena A Allen, Peter A Bandettini, Vince D Calhoun, Maurizio Corbetta, Stefania Della Penna, Jeff H Duyn, Gary H Glover, Javier Gonzalez-Castillo, Daniel A Handwerker, Shella Keilholz, Vesa Kiviniemi, David A Leopold, Francesco de Pasquale, Olaf Sporns, Martin Walter, and Catie Chang. 2013. Dynamic functional connectivity: Promise, issues, and interpretations. *NeuroImage* 80 (Oct. 2013), 360–378. DOI:http://dx.doi.org/10.1016/j.neuroimage.2013.05.079

Alon Itai and Michael Rodeh. 1978. Finding a Minimum Circuit in a Graph. *SIAM J. Comput.* 7, 4 (Nov. 1978), 413–423. DOI:http://dx.doi.org/10.1137/0207033

Richard James, Darren P. Croft, and Jens Krause. 2009. Potential banana skins in animal social network analysis. *Behavioral Ecology and Sociobiology* 63, 7 (2009), 989–997. DOI:http://dx.doi.org/10.1007/s00265-009-0742-5

Jaya Kawale, Snigdhansu Chatterjee, Dominick Ormsby, Karsten Steinhaeuser, Stefan Liess, and Vipin Kumar. 2012. Testing the Significance of Spatio-temporal Teleconnection Patterns. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. ACM, 642–650. DOI:http://dx.doi.org/10.1145/2339530.2339634

Jaya Kawale, Stefan Liess, Arjun Kumar, Michael Steinbach, Peter Snyder, Vipin Kumar, Auroop R Ganguly, Nagiza F Samatova, and Fredrick Semazzi. 2013. A graph-based approach to find teleconnections in climate data. *Statistical Analysis and Data Mining* 6, 3 (2013), 158–179. DOI:http://dx.doi.org/10.1002/sam.11181

Roland Kays, Margaret C Crofoot, Walter Jetz, and Martin Wikelski. 2015. Terrestrial animal tracking as an eye on life and planet. *Science* 348, 6240 (2015). DOI:http://dx.doi.org/10.1126/science.aaa2478

Matt J Keeling and Ken T D Eames. 2005. Networks and epidemic models. *Journal of The Royal Society Interface* 2, 4 (Sept. 2005), 295–307. DOI:http://dx.doi.org/10.1098/rsif.2005.0051

David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the Spread of Influence Through a Social Network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*. ACM, 137–146. DOI:http://dx.doi.org/10.1145/956750.956769

Angelika Kimmig, Lilyana Mihalkova, and Lise Getoor. 2014. Lifted graphical models: a survey. *Machine Learning* 99, 1 (2014), 1–45. DOI:http://dx.doi.org/10.1007/s10994-014-5443-2

Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. 2014. Multilayer networks. *Journal of Complex Networks* 2, 3 (2014), 203–271. DOI:http://dx.doi.org/10.1093/comnet/cnu016

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. 2014. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, 4 (2014), 795–816. DOI:http://dx.doi.org/10.1111/rssb.12050

Florian Klimm, Danielle S Bassett, Jean M Carlson, and Peter J Mucha. 2014. Resolving Structural Variability in Network Models and the Brain. *PLoS Comput Biol* 10, 3 (Mar. 2014), e1003491. DOI:http://dx.doi.org/10.1371/journal.pcbi.1003491

Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *Machine learning: ECML 2004*. Springer, 217–226. DOI:http://dx.doi.org/10.1007/978-3-540-30115-8_22

Eric D. Kolaczyk. 2009. Network Topology Inference. In *Statistical Analysis of Network Data SE - 7*. Springer New York, 1–48. DOI:http://dx.doi.org/10.1007/978-0-387-88146-1_7

Mark A Kramer, Uri T Eden, Sydney S Cash, and Eric D Kolaczyk. 2009. Network inference with confidence from multivariate time series. *Physical Review E* 79, 6 (Jun. 2009), 61916. DOI:http://dx.doi.org/10.1103/PhysRevE.79.061916

Mark A Kramer, Eric D Kolaczyk, and Heidi E Kirsch. 2008. Emergent network topology at seizure onset in humans. *Epilepsy Research* 79, 23 (2008), 173–186. DOI:http://dx.doi.org/10.1016/j.eplepsyres.2008.02.002

Jens Krause, Stefan Krause, Robert Arlinghaus, Ioannis Psorakis, Stephen Roberts, and Christian Rutz. 2013. Reality mining of animal social systems. *Trends in Ecology & Evolution* 28, 9 (Sept. 2013), 541–551. DOI:http://dx.doi.org/10.1016/j.tree.2013.06.002

Marlene Kretschmer, Dim Coumou, Jonathan F Donges, and Jakob Runge. 2016. Using Causal Effect Networks to analyze different Arctic drivers of mid-latitude winter circulation. *Journal of Climate* (Mar 2016). DOI:http://dx.doi.org/10.1175/JCLI-D-15-0654.1

Timothy La Fond and Jennifer Neville. 2010. Randomization Tests for Distinguishing Social Influence and Homophily Effects. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, 601–610. DOI:http://dx.doi.org/10.1145/1772690.1772752

Jean-Philippe Lachaux, Antoine Lutz, David Rudrauf, Diego Cosmelli, Michel Le Van Quyen, Jacques Martinerie, and Francisco Varela. 2002. Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence. *Neurophysiologie Clinique/Clinical Neurophysiology* 32, 3 (Jun. 2002), 157–

174. DOI:http://dx.doi.org/10.1016/S0987-7053(02)00301-5
N D Lane, E Miluzzo, H Lu, D Peebles, T Choudhury, and A T Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications Magazine* 48, 9 (Sept. 2010), 140–150. DOI:http://dx.doi.org/10.1109/MCOM.2010.5560598
Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Bornet, Trinh Minh Tri Do, Olivier Dousse, Julien Eberle, and Markus Miettinen. 2013. From big smartphone data to worldwide research: The Mobile Data Challenge. *Pervasive and Mobile Computing* 9, 6 (Dec. 2013), 752–771. DOI:http://dx.doi.org/10.1016/j.pmcj.2013.07.014
David Lazer. 2011. Networks in Political Science: Back to the Future. *PS: Political Science & Politics* 44, 01 (2011), 61–68. DOI:http://dx.doi.org/10.1017/S1049096510001873
Sophie Lebre, Jennifer Becq, Frederic Devaux, Michael Stumpf, and Gaelle Lelandais. 2010. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology* 4, 1 (2010), 130. DOI:http://dx.doi.org/10.1186/1752-0509-4-130
Jure Leskovec and Eric Horvitz. 2014. Geospatial Structure of a Planetary-Scale Social Network. *Computational Social Systems, IEEE Transactions on* 1, 3 (Sept. 2014), 156–163. DOI:http://dx.doi.org/10.1109/TCSS.2014.2377789
S Levitus, J I Antonov, O K Baranova, T P Boyer, C L Coleman, H E Garcia, A I Grodsky, D R Johnson, R A Locarnini, A V Mishonov, J R Reagan, C L Sazama, D Seidov, I Smolyar, E S Yarosh, and M M Zweng. 2013. The World Ocean Database. *Data Science Journal* 12 (2013), WDS229–WDS234. DOI:http://dx.doi.org/10.2481/dsj.WDS-041
H Li, J Xuan, Y Wang, and M Zhan. 2008. Inferring regulatory networks. *Frontiers in Bioscience* 13 (2008), 263–275. DOI:http://dx.doi.org/10.2741/2677
David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* 58, 7 (May 2007), 1019–1031. DOI:http://dx.doi.org/10.1002/asi.20591
Linyuan Lü and Tao Zhou. 2011. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* 390, 6 (2011), 1150–1170. DOI:http://dx.doi.org/10.1016/j.physa.2010.11.027
Subramani Mani and Gregory F Cooper. 2004. Causal discovery using a Bayesian local causal discovery algorithm. *Studies in Health Technology and Informatics* 107, Pt 1 (2004), 731–735. DOI:http://dx.doi.org/10.3233/978-1-60750-949-3-731
Daniel Marbach, James C Costello, Robert Kuffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, and Gustavo Stolovitzky. 2012. Wisdom of crowds for robust gene network inference. *Nature Methods* 9, 8 (Aug. 2012), 796–804. DOI:http://dx.doi.org/10.1038/nmeth.2016
Rossana Mastrandrea, Julie Fournet, and Alain Barrat. 2015. Contact Patterns in a High School: A Comparison between Data Collected Using Wearable Sensors, Contact Diaries and Friendship Surveys. *PLoS ONE* 10, 9 (Sept. 2015), e0136497. DOI:http://dx.doi.org/10.1371/journal.pone.0136497
Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring Networks of Substitutable and Complementary Products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, 785–794. DOI:http://dx.doi.org/10.1145/2783258.2783381
David B McDonald. 2007. Predicting fate from early connectivity in a social network. *Proceedings of the National Academy of Sciences* 104, 26 (Jun. 2007), 10910–10914. DOI:http://dx.doi.org/10.1073/pnas.0701159104
Christopher Meek. 1995. Causal Inference and Causal Explanation with Background Knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial*

Intelligence (UAI'95). Morgan Kaufmann Publishers Inc., 403–410. http://dl.acm.org/citation.cfm?id=2074204

Patrick Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. 2007. Information-Theoretic Inference of Large Transcriptional Regulatory Networks. *EURASIP Journal on Bioinformatics and Systems Biology* 2007, 1 (2007), 79879. DOI:http://dx.doi.org/10.1155/2007/79879

Lauren Ancel Meyers, Babak Pourbohloul, M E J Newman, Danuta M Skowronski, and Robert C Brunham. 2005. Network theory and SARS: predicting outbreak diversity. *Journal of Theoretical Biology* 232, 1 (2005), 71–81. DOI:http://dx.doi.org/10.1016/j.jtbi.2004.07.026

Isobel Milns, Colin M Beale, and V Anne Smith. 2010. Revealing ecological networks using Bayesian network inference algorithms. *Ecology* 91, 7 (Feb. 2010), 1892–1899. DOI:http://dx.doi.org/10.1890/09-0731.1

Sach Mukherjee and Terence P Speed. 2008. Network inference using informative priors. *Proceedings of the National Academy of Sciences* 105, 38 (Sept. 2008), 14313–14318. DOI:http://dx.doi.org/10.1073/pnas.0802272105

Seth Myers and Jure Leskovec. 2010. On the Convexity of Latent Social Network Inference. In *Advances in Neural Information Processing Systems 23*. 1741–1749. http://papers.nips.cc/paper/4113-on-the-convexity-of-latent-social-network-inference

Galileo Mark Namata, Ben London, and Lise Getoor. 2015. Collective Graph Identification. *ACM Transactions on Knowledge Discovery from Data* (2015). DOI:http://dx.doi.org/10.1145/2818378

G L Nemhauser, L A Wolsey, and M L Fisher. 1978. An analysis of approximations for maximizing submodular set functions–I. *Mathematical Programming* 14, 1 (1978), 265–294. DOI:http://dx.doi.org/10.1007/BF01588971

Praneeth Netrapalli and Sujay Sanghavi. 2012. Learning the Graph of Epidemic Cascades. *SIGMETRICS Perform. Eval. Rev.* 40, 1 (Jun. 2012), 211–222. DOI:http://dx.doi.org/10.1145/2318857.2254783

Jian Ni, Haiyong Xie, Sekhar Tatikonda, and Yang Richard Yang. 2010. Efficient and Dynamic Routing Topology Inference from End-to-end Measurements. *IEEE/ACM Transactions on Networking* 18, 1 (Feb. 2010), 123–135. DOI:http://dx.doi.org/10.1109/TNET.2009.2022538

M Paluš, D Hartman, J Hlinka, and M Vejmelka. 2011. Discerning connectivity from dynamics in climate networks. *Nonlinear Processes in Geophysics* 18, 5 (2011), 751–763. DOI:http://dx.doi.org/10.5194/npg-18-751-2011

Evangelos E Papalexakis, Alona Fyshe, Nicholas D Sidiropoulos, Partha Pratim Talukdar, Tom M Mitchell, and Christos Faloutsos. 2014. Good-enough Brain Model: Challenges, Algorithms and Discoveries in Multi-subject Experiments. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, 95–104. DOI:http://dx.doi.org/10.1145/2623330.2623639

D M Patrick, M L Rekart, A Jolly, S Mak, M Tyndall, J Maginley, E Wong, T Wong, H Jones, C Montgomery, and R C Brunham. 2002. Heterosexual outbreak of infectious syphilis: epidemiological and ethnographic analysis and implications for control. *Sexually Transmitted Infections* 78, suppl 1 (2002), i164–i169. DOI:http://dx.doi.org/10.1136/sti.78.suppl_1.i164

G Pfurtscheller and C Andrew. 1999. Event-Related changes of band power and coherence: methodology and interpretation. *Journal of Clinical Neurophysiology* 16, 6 (Nov. 1999), 512–519. http://dx.doi.org/10.1097/00004691-199911000-00003

Noa Pinter-Wollman, Elizabeth A Hobson, Jennifer E Smith, Andrew J Edelman, Daizaburo Shizuka, Shermin de Silva, James S Waters, Steven D Prager, Takao Sasaki, George Wittemyer, Jennifer Fewell, and David B McDonald. 2013. The dynamics of animal social networks: analytical, conceptual, and theoretical advances. *Behavioral*

Ecology (2013). DOI:http://dx.doi.org/10.1093/beheco/art047

S C Ponten, F Bartolomei, and C J Stam. 2016. Small-world networks and epilepsy: Graph theoretical analysis of intracerebrally recorded mesial temporal lobe seizures. *Clinical Neurophysiology* 118, 4 (Apr. 2016), 918–927. DOI:http://dx.doi.org/10.1016/j.clinph.2006.12.002

Stephen R Proulx, Daniel E L Promislow, and Patrick C Phillips. 2005. Network thinking in ecology and evolution. *Trends in Ecology & Evolution* 20, 6 (2005), 345–353. DOI:http://dx.doi.org/10.1016/j.tree.2005.04.004

N Pržulj, D G Corneil, and I Jurisica. 2004. Modeling interactome: scale-free or geometric? *Bioinformatics* 20, 18 (2004), 3508–3515. DOI:http://dx.doi.org/10.1093/bioinformatics/bth436

Ioannis Psorakis, Stephen J Roberts, Iead Rezek, and Ben C Sheldon. 2012. Inferring social network structure in ecological systems from spatio-temporal data streams. *Journal of The Royal Society Interface* 9, 76 (2012), 3055–3066. DOI:http://dx.doi.org/10.1098/rsif.2012.0223

Manish Purohit, B Aditya Prakash, Chanhyun Kang, Yao Zhang, and V S Subrahmanian. 2014. Fast Influence-based Coarsening for Large Networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, 1296–1305. DOI:http://dx.doi.org/10.1145/2623330.2623701

J D Ramsey, S J Hanson, C Hanson, Y O Halchenko, R A Poldrack, and C Glymour. 2010. Six problems for causal inference from fMRI. *NeuroImage* 49, 2 (2010), 1545–1558. DOI:http://dx.doi.org/10.1016/j.neuroimage.2009.08.065

Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H Strogatz. 2010. Redrawing the Map of Great Britain from a Network of Human Interactions. *PLoS ONE* 5, 12 (Dec. 2010), e14248. DOI:http://dx.doi.org/10.1371/journal.pone.0014248

Jaap C Reijneveld, Sophie C Ponten, Henk W Berendse, and Cornelis J Stam. 2007. The application of graph theoretical analysis to complex networks in the brain. *Clinical Neurophysiology* 118, 11 (Dec. 2007), 2317–2331. DOI:http://dx.doi.org/10.1016/j.clinph.2007.08.010

Lucy F Robinson, Lauren Y Atlas, and Tor D Wager. 2015. Dynamic functional connectivity using state-based dynamic community structure: Method and application to opioid analgesia. *NeuroImage* 108 (2015), 274–291. DOI:http://dx.doi.org/10.1016/j.neuroimage.2014.12.034

Alard Roebroeck, Elia Formisano, and Rainer Goebel. 2005. Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage* 25, 1 (2005), 230–242. DOI:http://dx.doi.org/10.1016/j.neuroimage.2004.11.017

M J Rosa, K Friston, and W Penny. 2012. Post-hoc selection of dynamic causal models. *Journal of Neuroscience Methods* 208, 1 (2012), 66–78. DOI:http://dx.doi.org/10.1016/j.jneumeth.2012.04.013

Mikail Rubinov and Olaf Sporns. 2010. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* 52, 3 (Sept. 2010), 1059–1069. DOI:http://dx.doi.org/10.1016/j.neuroimage.2009.10.003

Jakob Runge, Vladimir Petoukhov, and Jürgen Kurths. 2013. Quantifying the Strength and Delay of Climatic Interactions: The Ambiguities of Cross Correlation and a Novel Measure Based on Graphical Models. *Journal of Climate* 27, 2 (Sept. 2013), 720–739. DOI:http://dx.doi.org/10.1175/JCLI-D-13-00159.1

Christian Rutz, Zackory T Burns, Richard James, Stefanie M H Ismar, John Burt, Brian Otis, Jayson Bowen, and James J H St Clair. 2012. Automated mapping of social networks in wild birds. *Current Biology* 22, 17 (2012), R669 – R671. DOI:http://dx.doi.org/10.1016/j.cub.2012.06.037

Suranjana Saha, Shrinivas Moorthi, Xingren Wu, Jiande Wang, Sudhir Nadiga, Patrick Tripp, David Behringer, Yu-Tai Hou, Hui-ya Chuang, Mark Iredell, Michael Ek, Jesse Meng, Rongqian Yang, Malaquías Peña Mendez, Huug van den Dool, Qin Zhang, Wanqiu Wang, Mingyue Chen, and Emily Becker. 2014. The NCEP Climate Forecast System Version 2. *Journal of Climate* 27, 6 (2014), 2185–2208. DOI:http://dx.doi.org/10.1175/JCLI-D-12-00823.1

V Sakkalis. 2011. Review of Advanced Techniques for the Estimation of Brain Connectivity Measured with EEG/MEG. *Computers in Biology and Medicine* 41, 12 (Dec. 2011), 1110–1117. DOI:http://dx.doi.org/10.1016/j.compbiomed.2011.06.020

Jari Saramäki and Esteban Moro. 2015. From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *The European Physical Journal B* 88, 6 (2015), 1–10. DOI:http://dx.doi.org/10.1140/epjb/e2015-60106-6

Vedran Sekara and Sune Lehmann. 2014. The Strength of Friendship Ties in Proximity Sensor Data. *PLoS ONE* 9, 7 (July 2014), e100915. DOI:http://dx.doi.org/10.1371/journal.pone.0100915

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective Classification in Network Data. *AI Magazine* 29, 3 (2008), 93. http://hdl.handle.net/1903/7546

Jay Shendure and Hanlee Ji. 2008. Next-generation DNA sequencing. *Nature Biotechnology* 26, 10 (Oct. 2008), 1135–1145. DOI:http://dx.doi.org/10.1038/nbt1486

Chao Sima, Jianping Hua, and Sungwon Jung. 2009. Inference of gene regulatory networks using time-series data: a survey. *Current Genomics* 10, 6 (Sept. 2009), 416–429. DOI:http://dx.doi.org/10.2174/138920209789177610

Sean L Simpson, F DuBois Bowman, and Paul J Laurienti. 2013. Analyzing complex functional brain networks: Fusing statistics and network science to understand the brain. *Statistics Surveys* 7 (2013), 1–36. DOI:http://dx.doi.org/10.1214/13-SS103

Timo Smieszek, Stefanie Castell, Alain Barrat, Ciro Cattuto, Peter J. White, and G'erard Krause. 2016. Contact Diaries Versus Wearable Proximity Sensors in Measuring Contact Patterns at a Conference: Method Comparison and Participants' Attitudes. *BMC Infectious Diseases* 16, 1 (2016), 1–14. DOI:http://dx.doi.org/10.1186/s12879-016-1676-y

Olaf Sporns. 2014. Contributions and challenges for network models in cognitive neuroscience. *Nature Neuroscience* 17, 5 (May 2014), 652–660. DOI:http://dx.doi.org/10.1038/nn.3690

Olaf Sporns and Richard F Betzel. 2016. Modular Brain Networks. *Annual Review of Psychology* 67, 1 (Jan. 2016), 613–640. DOI:http://dx.doi.org/10.1146/annurev-psych-122414-033634

J Conrad Stack, Shweta Bansal, V S Anil Kumar, and Bryan Grenfell. 2013. Inferring population-level contact heterogeneity from common epidemic data. *Journal of The Royal Society Interface* 10, 78 (2013). DOI:http://dx.doi.org/10.1098/rsif.2012.0578

Juliette Stehl, Franois Charbonnier, Tristan Picard, Ciro Cattuto, and Alain Barrat. 2013. Gender Homophily from Spatial Behavior in a Primary School: A Sociometric Study. *Social Networks* 35, 4 (2013), 604 – 613. DOI:http://dx.doi.org/10.1016/j.socnet.2013.08.003

Karsten Steinhaeuser, Nitesh V Chawla, and Auroop R Ganguly. 2011. Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. *Statistical Analysis and Data Mining* 4, 5 (2011), 497–511. DOI:http://dx.doi.org/10.1002/sam.10100

A Stopczynski, R Pietri, A Pentland, D Lazer, and S Lehmann. 2014a. Privacy in Sensor-Driven Human Data Collection: A Guide for Practitioners. *ArXiv e-prints* (Mar. 2014). https://arxiv.org/abs/1403.5299

Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My

Madsen, Jakob Eg Larsen, and Sune Lehmann. 2014b. Measuring Large-Scale Social Networks with High Resolution. *PLoS ONE* 9, 4 (Apr. 2014), e95978. DOI:http://dx.doi.org/10.1371/journal.pone.0095978

Cédric Sueur, Armand Jacobs, Frédéric Amblard, Odile Petit, and Andrew J King. 2011. How can social network analysis improve the study of primate behavior? *American Journal of Primatology* 73, 8 (Aug. 2011), 703–19. DOI:http://dx.doi.org/10.1002/ajp.20915

Kaustubh Supekar, Vinod Menon, Daniel Rubin, Mark Musen, and Michael D Greicius. 2008a. Network Analysis of Intrinsic Functional Brain Connectivity in Alzheimer's Disease. *PLoS Comput Biol* 4, 6 (2008), 1–11. DOI:http://dx.doi.org/10.1371/journal.pcbi.1000100

Kaustubh Supekar, Vinod Menon, Daniel Rubin, Mark Musen, and Michael D Greicius. 2008b. Network Analysis of Intrinsic Functional Brain Connectivity in Alzheimer's Disease. *PLoS Computational Biology* 4, 6 (Jun. 2008), e1000100. DOI:http://dx.doi.org/10.1371/journal.pcbi.1000100

Hiroyuki Toh and Katsuhisa Horimoto. 2002. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics* 18, 2 (Feb. 2002), 287–297. DOI:http://dx.doi.org/10.1093/bioinformatics/18.2.287

A A Tsonis and P J Roebber. 2004. The architecture of the climate network. *Physica A: Statistical Mechanics and its Applications* 333, 0 (2004), 497–504. DOI:http://dx.doi.org/10.1016/j.physa.2003.10.045

Anastasios A. Tsonis, Geli Wang, Kyle L. Swanson, Francisco A. Rodrigues, and Luciano da Fontura Costa. 2011. Community structure and dynamics in climate networks. *Climate Dynamics* 37, 5-6 (2011), 933–940. DOI:http://dx.doi.org/10.1007/s00382-010-0874-3

Charalampos E Tsourakakis, U Kang, Gary L Miller, and Christos Faloutsos. 2009. DOULION: Counting Triangles in Massive Graphs with a Coin. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. ACM, 837–846. DOI:http://dx.doi.org/10.1145/1557019.1557111

Martijn P van den Heuvel, René S Kahn, Joaquín Goñi, and Olaf Sporns. 2012. High-cost, high-capacity backbone for global brain communication. *Proceedings of the National Academy of Sciences* 109, 28 (July 2012), 11372–11377. DOI:http://dx.doi.org/10.1073/pnas.1203593109

Martijn P van den Heuvel and Olaf Sporns. 2011. Rich-club organization of the human connectome. *The Journal of Neuroscience* 31, 44 (Nov. 2011), 15775–15786. DOI:http://dx.doi.org/10.1523/jneurosci.3539-11.2011

Martijn P van den Heuvel and Olaf Sporns. 2015. Network hubs in the human brain. *Trends in Cognitive Sciences* 17, 12 (Dec. 2015), 683–696. DOI:http://dx.doi.org/10.1016/j.tics.2013.09.012

Dan J Wang, Xiaolin Shi, Daniel A McFarland, and Jure Leskovec. 2012. Measurement error in network data: A re-classification. *Social Networks* 34, 4 (2012), 396–409. DOI:http://dx.doi.org/10.1016/j.socnet.2012.01.003

Larry Wasserman. 2006. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer-Verlag New York, Inc.

Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* 393, 6684 (Jun. 1998), 440–442. DOI:http://dx.doi.org/10.1038/30918

David Welch, Shweta Bansal, and David R Hunter. 2011. Statistical inference to advance network models in epidemiology. *Epidemics* 3, 1 (Mar. 2011), 38–45. DOI:http://dx.doi.org/10.1016/j.epidem.2011.01.002

Tina Wey, Daniel T Blumstein, Weiwei Shen, and Ferenc Jordán. 2008. Social network analysis of animal behaviour: a promising tool for the study of sociality. *Animal Behaviour* 75, 2 (2008), 333–344. DOI:http://dx.doi.org/10.1016/j.anbehav.2007.06.020

Hal Whitehead, Lars Bejder, and C Andrea Ottensmeyer. 2005. Testing association patterns: issues arising and extensions. *Animal Behaviour* 69, 5 (2005), e1 –. DOI:http://dx.doi.org/10.1016/j.anbehav.2004.11.004

Hal Whitehead and Richard James. 2015. Generalized affiliation indices extract affiliations from social network data. *Methods in Ecology and Evolution* 6, 7 (2015), 836–844. DOI:http://dx.doi.org/10.1111/2041-210X.12383

R W Wilkins, D A Hodges, P J Laurienti, M Steen, and J H Burdette. 2014. Network Science and the Effects of Music Preference on Functional Brain Connectivity: From Beethoven to Eminem. *Scientific Reports* 4 (Aug. 2014), 6130. DOI:http://dx.doi.org/10.1038/srep06130

Rongjing Xiang and Jennifer Neville. 2008. Pseudolikelihood EM for Within-network Relational Learning. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008)*. 1103–1108. DOI:http://dx.doi.org/10.1109/ICDM.2008.148

K Yamasaki, A Gozolchiani, and S Havlin. 2008. Climate Networks around the Globe are Significantly Affected by El Niño. *Phys. Rev. Lett.* 100, 22 (Jun. 2008), 228501. DOI:http://dx.doi.org/10.1103/PhysRevLett.100.228501

Donghyeon Yu, MinSoo Kim, Guanghua Xiao, and Tae Hyun Hwang. 2013. Review of Biological Network Data and Its Applications. *Genomics & Informatics* 11, 4 (Dec. 2013), 200–210. DOI:http://dx.doi.org/10.5808/GI.2013.11.4.200

Qingbao Yu, Erik B Erhardt, Jing Sui, Yuhui Du, Hao He, Devon Hjelm, Mustafa S Cetin, Srinivas Rachakonda, Robyn L Miller, Godfrey Pearlson, and Vince D Calhoun. 2015. Assessing dynamic brain graphs of time-varying connectivity in fMRI data: Application to healthy controls and patients with schizophrenia. *NeuroImage* 107 (2015), 345–355. DOI:http://dx.doi.org/10.1016/j.neuroimage.2014.12.020

Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68, 1 (Feb. 2006), 49–67. DOI:http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x

Andrew Zalesky, Alex Fornito, and Ed Bullmore. 2012. On the use of correlation as a measure of network connectivity. *NeuroImage* 60, 4 (2012), 2096–2106. DOI:http://dx.doi.org/10.1016/j.neuroimage.2012.02.001

Andrew Zalesky, Alex Fornito, Ian H Harding, Luca Cocchi, Murat Yücel, Christos Pantelis, and Edward T Bullmore. 2010. Whole-brain anatomical networks: Does the choice of nodes matter? *NeuroImage* 50, 3 (Apr. 2010), 970–983. DOI:http://dx.doi.org/10.1016/j.neuroimage.2009.12.027

Tanja Zerenner, Petra Friederichs, Klaus Lehnertz, and Andreas Hense. 2014. A Gaussian graphical model approach to climate networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 24, 2 (2014). DOI:http://dx.doi.org/10.1063/1.4870402

Yang Zhan, David Halliday, Ping Jiang, Xuguang Liu, and Jianfeng Feng. 2006. Detecting time-dependent coherence between non-stationary electrophysiological signalsA combined statistical and timefrequency approach. *Journal of Neuroscience Methods* 156, 12 (Sept. 2006), 322–332. DOI:http://dx.doi.org/10.1016/j.jneumeth.2006.02.013

Bin Zhang and Steve Horvath. 2005. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology* 4, 1 (2005). DOI:http://dx.doi.org/10.2202/1544-6115.1128

Dong Zhou, Avi Gozolchiani, Yosef Ashkenazy, and Shlomo Havlin. 2015. Teleconnection Paths via Climate Network Direct Link Detection. *Phys. Rev. Lett.* 115, 26 (Dec. 2015), 268501. DOI:http://dx.doi.org/10.1103/PhysRevLett.115.268501

Hui Zhou, Wencai Du, Shaochun Xu, and Qinling Xin. 2011. An Empirical Study of Network Topology Inference. In *Computer and Information Science 2011 SE - 17*,

Roger Lee (Ed.). Studies in Computational Intelligence, Vol. 364. Springer Berlin Heidelberg, 213–225. DOI: http://dx.doi.org/10.1007/978-3-642-21378-6_17