Contents

Chapter 2

Structural Properties of Networks over Time

A fundamental question in dynamic network analysis is to determine how the graph-theoretic properties of a network are changing over time when the network is not completely visible. A number of diverse questions depend on our ability to accurately measure structural changes in real networks: for example, is the average number of hops between Internet routers increasing or decreasing as the Internet grows, and does the pattern of collaborations between publishing scientists indicate that scientists are forming an increasingly tighter-connected social network? These issues are important in computer network analysis as well as physics and sociology, and in general, any domain where a physical system can be represented as a partially observable *graph structured process* that generates a changing network over time. In these cases, the change in the structure of the physical system between any two time points is quantified by the change in some graph-theoretic measure evaluated at those points.

There are at least two reasons to look for trends in the time evolution of various graph measures.

- *Generative network models.* Since the true dynamical processes that generate real dynamic network datasets are too complex to model, a line of research aims to develop simple dynamical processes that can approximate the true process by generating networks with a given trend over time in some graph measure [Leskovec et al., 2007, Akoglu et al., 2008, Akoglu and Faloutsos, 2009, Bonato et al., 2009, Chakrabarti et al., 2010, Du et al., 2010]. These dynamical processes can very generally be viewed as probabilistic algorithms driven by uniform random noise that convert one graph into another. Inferring the parameters of these processes from data is generally intractable, so a common method of justifying various dynamical graph generation models is to show that they reproduce temporal trends in measures comparable to an average temporal trend found in multiple real datasets. Naturally, this presumes that the temporal trends measured from empirical data were accurate and not artifacts of the data sampling process, an issue that we show is not trivial.

- *Scientific datasets.* In scientific datasets, a sudden fundamental change in the structure of a network can sometimes be detected by tracking a correspondingly large change in an appropriate graph measure. In animal association networks from ecology, for example, differences in network measures are used to

determine behavioral differences between species [Sundaresan et al., 2007].

In an ideal world, we would simply obtain a sequence of snapshots of a graph structured process over time and periodically measure properties of interest in order to determine a trend. In the real world, however, our observational capabilities are severely limited in terms of how far back in time datasets reach, how complete each snapshot is, and the mechanisms by which network data is collected. In this chapter, we examine how these issues are tackled by surveying the growing literature on dynamic network measurements. Several recent empirical studies suggest that the structures of a variety of otherwise disparate real networks are evolving in similar ways [Leskovec et al., 2005, Ahn et al., 2007, Menezes et al., 2009]. We describe in detail the experimental methodology used to reach these conclusions, and investigate how sensitive they are to various kinds of missing data.

This chapter is organized as follows. In the next section, we present a methodology survey describing typical procedures for measuring graphical properties of a network over time, as well as a characterization of dynamic network datasets into two broad classes: *interaction* networks, and *citation* networks, depending on the nature and permanence of interactions (edges). Section 2.2 presents a literature review and summary of published empirical results on the graphical properties of real dynamic networks over time. All the published results on networks that we survey can be classified as either interaction or citation networks. In order to assess the significance of the many common temporal trends in network properties (*e.g.,* , decreasing average shortest path length over time), we experimentally analyze whether these trends necessarily reflect the reality of the underlying system in the presence of missing data and observational limitations in the data collection process. Section 2.3 deals with citation networks, and Section 2.4 deals with the more common class of interaction networks. In Section 2.5, we summarize our findings and list some suggestions for future research.

2.1 Networks, measurement, and error

Dynamic networks are explicit representations of the change in the structure of a physical system over time. These systems are generally complex real-world phenomena that can be modeled as a set of uniquely identifiable entities interacting with each other over time. The entities can act independently of each other, and the interactions between them can occur in arbitrary ways. The global structure of which entities have interacted with which other entities can therefore be a complex network-like structure that (presumably) evolves over time. Quantifying the change in the structure of such a system over time is the first step in analyzing a dynamic network.

The most immediate way of quantifying the change in a dynamic network between two time points is to simply measure its structure at both time points and compute the difference. There are numerous ways to measure a graph, but in a very general sense that covers the most popular methods, a measure M can be thought of as a function from a graph to a real number.

$$M : G \mapsto \mathbb{R} \quad \text{where } G = (V, E)$$

Some simple measures include the number of nodes and edges in a graph, the average shortest path length between all pairs of connected nodes, and the largest eigenvalue of the adjacency matrix. These are *static*

measures that operate on a single graph; an example of a measure that operates on a graph through time is the graph edit distance measure between consecutive graphs used in anomaly detection in computer networks [Bunke, 2000]. Given a dynamic network dataset, one could simply construct a graph of all interactions up to a given time point, measure the graph, and subsequently produce a time series of measure values by repeating the process, which would then presumably characterize the change in the structure of the system.

A number of assumptions need to hold before this is true. Clearly, the measure M that is used must adequately represent true change in the physical system, at least with respect to the reason for measuring the dynamic network. For example, if we want to quantify the change in a computer network to study the effect of decentralized routing protocols, the average shortest path length might be a good measure, whereas the raw number of nodes might not be. Secondly, if there are sampling and incomplete data issues in the dataset (as is the case with almost all empirical time-series datasets), then the chosen measure must be robust to sampling errors. In particular, if we are to measure a dynamic network at successive time points in the presence of sampling error, then the resultant time series of measure values should be similar to the true underlying time series, at least qualitatively.

There also exist two common methods for aggregating dynamic network data into a single graph that can be measured. For example, given a dynamic network dataset, we might have identified two time points at which we wish to measure the structure of the underlying system. Both aggregation methods can be applied to the same data under certain conditions, and yield very different results because they make different assumptions about the data. Adding to the subtleties of choosing an aggregation method, we have also identified two different classes of network data that require different treatment. This distinction between aggregation methods and network data classes does not appear to have been explicitly made, particularly in terms of measurement biases involved in mismatching aggregation methods, graph measures, and network data classes. We start this analysis with a description of network data classes in the next subsection, and network aggregation methods in Section 2.1.2.

2.1.1 Network data classes

All the dynamic network datasets described in the literature that we will later review in Section 2.2 fall into one of two categories depending on the type of physical system being modeled: *interaction* networks and *citation* networks. The overall characteristics of each class are as follows.

- *(Interaction networks)* An edge, identified by the labels of its adjacent nodes, can appear at multiple timesteps. Examples include email networks built from email headers [Leskovec et al., 2005, Diesner and Carley, 2005], where people continually send emails to each other over time, and similar networks built from phone records [Nanavati et al., 2006], logs from physical proximity sensors [Eubank et al., 2004, Eagle and Pentland, 2006], and co-authorship of published scientific articles [Newman, 2001c, Barabási et al., 2002, Leskovec et al., 2005]. Each edge in the dynamic network dataset is therefore a record of a single instance of an interaction between nodes, out of many other possible instances at different times. Network structure is determined through the proxy of interactions, which can be regulated by an independent dynamic process.

- *(Citation networks)* Unlike interaction networks, citation networks only grow in size over time with the

addition of new nodes and edges, and each edge appears only once in the observation stream. Thus, in a dynamic network dataset, the appearance of an edge at time t implies that the underlying physical system grew by at least one edge at time t. Their name comes from the canonical example of the directed network of citations between scientific papers; once a scientific paper is published, its citations never change, and thus each edge in the dynamic network appears only once over all of time [Bilke and Peterson, 2001, Nerur et al., 2005, Leskovec et al., 2007]. Note that interaction networks where edges and nodes are deleted very infrequently relative to growth can also be seen as citation networks at short times. This covers some types of online social network data where the rate of node and edge addition far outstrips its removal, to the point where the online 'friendship' networks can be seen as citation networks [Kumar et al., 2006,Krishnamurthy et al., 2008]. For the same reason, it also includes domains such as blog inter-linking networks [Adar and Adamic, 2005].

The distinction between the two classes is important when we use network measures to represent the change in the underlying physical system. In particular, when a citation network dataset contains a record of edge (u, v) at time t, we know that the underlying physical system has increased in size by at least one edge. However, with interaction networks, particularly when we do not have the entire temporal history of the network, we cannot tell if the record of edge (u, v) at time t represents true growth in the underlying system, or is just the re-occurrence of an interaction that had also occurred some time before. In other words, there is a process that generates interactions along a growing network structure, and using interactions as a proxy for network structure requires that we account for the sampling error induced by the dynamics of the interactions, particularly at short times and without full temporal history. This is closely tied to the aggregation method (or lack thereof) used to assemble interaction data into a time series of graphs, which is the topic of the next subsection.

2.1.2 Network aggregation methods

We can view *citation* and *interaction* network data types through the lens of a sampling process (generally involved in data collection) on a true underlying network that is growing over time. The sampling process generates a sequence of graphs drawn from the true underlying process, which we call the *observations*. In the case of interaction networks, the sampling process reveals a set of interactions each time a group of entities interacts. In the case of citation networks, observations consist solely of (subsets of) new edges and vertices that are added to the network, with the constraint that the same edge must not appear twice. These samples are generally revealed over time, starting from an arbitrary time relative to the underlying process, from which we must infer both an initial structure as well as any change in it.

The earliest approach to network data was to treat the underlying process as being in a *steady state* with respect to some graph measure M.

Definition 2.1.1. *(Steady state)* A physical system is in equilibrium or a *steady state* with respect to a graph theoretic measure M if the partial derivative of M measured on the system with respect to time is zero.

$$\frac{\partial M}{\partial t} = 0$$

Vertices and edges may therefore be added to or deleted from the system, but the dynamic behavior of a

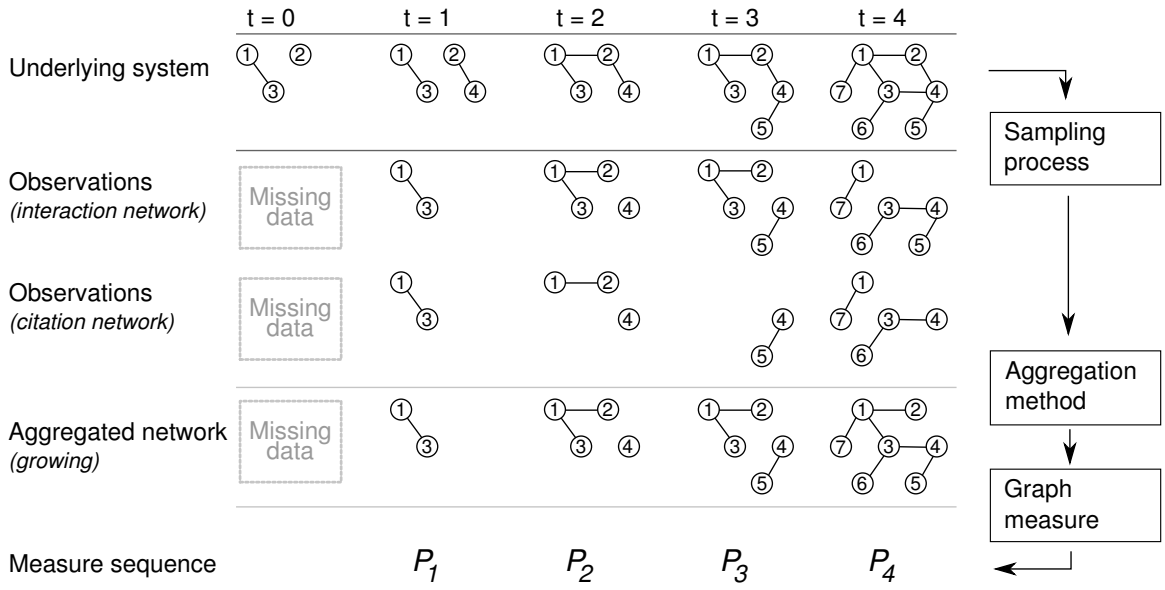


Figure 2.1: Example of a true underlying network, a sequence of observations of it starting at $t = 1$, and growing and dynamic interaction representations of the observations. Edge thickness represents edge weight.

measure of interest M is independent of time.[1] If we assume that an underlying system is in a steady state with respect to measure M, and assuming a consistent sampling process, we can simply let the observation sequence converge to a large enough graph and assume that M measured on this graph is representative of the true value of M. These steady state and consistency assumptions are the foundation of *static network analysis*, which approximates partially stationary dynamic processes as large networks. Static network analysis has produced a large number of important advances in network theory, such as the observations that real-world networks tend to have heavily skewed degree distributions and low average pairwise shortest path lengths [Albert and Barabási, 2002, Newman, 2003, Boccaletti et al., 2006]. A number of studies also analyze dynamical aspects of the sampling process, such as how long typical systems must be observed till various graph theoretic properties converge to limiting values [Latapy and Magnien, 2006], and how sensitive measured properties are to different types of sampling error [Costenbader and Valente, 2003,Kossinets, 2006].

There has also been interest in the dynamic behavior of some measure M of the underlying system, with the implicit assumption that the underlying system is not in a steady state with respect to M [Barabási et al., 2002, Leskovec et al., 2005]. Intuitively, the steady state assumption might not appear to be appropriate for systems as large and chaotic as the Internet (for example). In this case, observations are aggregated over time, either as a *growing* network or a *fully dynamic network*, into a single graph that is measured with M at fixed time intervals. Figure 2.1 depicts how a sampling process on the true underlying network generates observations, depending on whether the underlying network is a citation or interaction network. These samples of network structure are aggregated using the growing network method to yield a time series of graphs with measure values $P_1, ..., P_n$. We describe each aggregation method in more detail below.

Definition 2.1.2. *(Growing network)* Observations of the physical system are made at discrete timesteps.

[1]For example, a growing k-regular graph, one one in which every vertex has exactly k adjacent edges, is in equilibrium with respect to its degree distribution or any statistic of it, but not necessarily with respect to other measures.

Each observation is accumulated into a single growing graph, which is measured at fixed time intervals. Aggregating in this manner implicitly assumes that the underlying process can be represented as a citation network. Let $\mathbf{G}$ be the true, underlying network of the physical system being modeled, which is only partially observable and growing over time with the addition of new vertices and edges (by definition, vertices and edges are never removed in a citation network). Since $\mathbf{G}$ is growing over time, we let its structure at time t be denoted by an element in the sequence $\langle \mathbf{G} \rangle$.

$$\langle \mathbf{G} \rangle = \langle \mathbf{G}_0, ... \rangle \quad \text{where } \mathbf{G}_t = (\mathbf{V_t}, \mathbf{E_t})$$

where $\mathbf{G}_t$ is the *complete* network structure at time t, consisting of a set of labeled vertices $\mathbf{V_t} = \{\mathbf{v}_i : \mathbf{v}_i \in \mathbb{N}\}$ and a set of directed or undirected edges $\mathbf{E_t} = \{(\mathbf{v}_i, \mathbf{v}_j) : \mathbf{v}_i, \mathbf{v}_j \in \mathbf{V}_t\}$. Since vertices and edges are only added to the system, we have a monotone property on the vertex and edge sets over time.

$$\begin{aligned} \mathbf{V_t} &\subseteq \mathbf{V_{t+1}} \\ \mathbf{E_t} &\subseteq \mathbf{E_{t+1}} \end{aligned} \tag{2.1}$$

Since the true underlying network is only partially observable, we instead obtain finite samples of its graph structure over time, called the observation sequence $\langle O \rangle$, starting at some arbitrary time $t_0 > 0$.

$$\langle O \rangle = \langle G_{t_0}, ..., G_{t_n} \rangle \quad \text{where } G_{t_i} = (V_{t_i}, E_{t_i}),\ V_{t_i} \subseteq \mathbf{V}_{t_i}, E_{t_i} \subseteq \mathbf{E}_{t_i}$$

Due to limitations of the observation process, each observed graph G_t may be a small fraction of the size of the physical system at that point in time $\mathbf{G_t}$, *i.e.,* $|V_t| \ll |\mathbf{V_t}|$ and $|E_t| \ll |\mathbf{E_t}|$. The growing network aggregation method approximates $\mathbf{G_t}$ by aggregating all samples from the start of observations t_0 up to the current timestep into a single growing network $\langle G^+ \rangle$:

$$\langle G^+ \rangle = \langle G^+_{t_0}, ..., G^+_{t_n} \rangle \quad \text{where } G^+_t = (\bigcup_{i=t_0}^{t_n} V_i, \bigcup_{i=t_0}^{t_n} E_i)$$

Any graph measure M can then be measured on G^+_t at fixed intervals, yielding a time-series that (presumably) approximates the trend of the same property on the underlying network $\langle \mathbf{G} \rangle$.

A key assumption in the growing network methodology in practice is that the aggregated growing network at any point in time is a good approximation of the underlying physical system, including at the (necessarily) arbitrarily chosen time of the first observation t_0. Particularly, the assumption that we can accurately track trends in the underlying process, either quantitatively or qualitatively, by measuring an aggregated graph of observations periodically requires that the measured trend accurately track the trend in the underlying system.

Definition 2.1.3. *(Fully dynamic network)* A fully dynamic network approach treats the underlying system as an interaction network. Observations are aggregated within successive, non-overlapping intervals of time, and the graph obtained from each interval is measured independently of the other intervals. When the interval of time is fixed, it is called the *window* of the aggregation. It can be seen as a growing network where vertices and edges are removed after a fixed period of time (the window) unless they appear again

within the window. Thus, depending on the relative rates of vertex and edge addition and removal, a dynamic network may be structurally growing, in equilibrium, or shrinking. The underlying network $\langle \mathbf{G} \rangle$ no longer necessarily has the monotonicity property of growing networks (Equation 2.1).

A variant of this approach weights each edge with a time-decaying function, to emphasize the impact of more recent changes in graph structure. This is usually practiced in one of two related ways:

1. Weight each edge by the amount of time that has elapsed since it was last seen, and use algorithms that take the weight of each edge into account (*e.g.,* [Sharan and Neville, 2007, Acar et al., 2009, Barrat et al., 2004]).

2. Remove edges from the network after a certain period of time if they have not appeared again, where the amount of time is determined by a decay function (*e.g.,* [Kossinets and Watts, 2006]).

The issue of time-decaying relationships in social networks has also received attention in sociology, such as the formulation of sociologically meaningful decay functions suggested by data [Burt, 2000].

A practical issue with the fully dynamic aggregation method is that it presumes knowledge of an appropriate window length to aggregate observations over. In practice, this is usually determined by taking some 'natural' quantization of the dataset, such as a window length of a year for scientific publication data, but this is difficult to determine for some datasets (*e.g.,* animal association data). Furthermore, some measures M might be sensitive to the window length, in effect showing trends that are a function of the quantization window. However, more rigorous methods are needed to determine a good aggregation window for a dataset, and a complementary set of approaches attempts to find a time aggregation that minimizes some notion of error in measured dynamic trends [Sun et al., 2007, Sulo et al., 2010]. The growing network method has the practical advantage of not requiring any additional parameters, and we focus more on this method for the remainder of this chapter.

To illustrate the different situations that might call for growing networks over dynamic networks, consider the analysis of e-mail transmission records to deduce the structure of the underlying association network [Shetty and Adibi, 2005, Kossinets and Watts, 2006, Leskovec et al., 2007]. Treating each day as a timestep, the observation corresponds to the graph structure of e-mails that are sent and received on a given day. This has definite meaning, as we can investigate the time stream for periodic or other temporal patterns [Lahiri and Berger-Wolf, 2008, Lahiri and Berger-Wolf, 2010, Sun et al., 2007, Acar et al., 2009]. On the other hand, consider the discovery of Autonomous Systems (AS) routes on the Internet through the use of `traceroute` probes and similar methods [Andersen et al., 2002, Chen et al., 2002, Vázquez et al., 2002, Leskovec et al., 2007]. The temporal sequence of samples is a product of the observation technique and has no inherent meaning to the object of interest (the AS topology). Graph theoretic analysis of the sampled graph would be more meaningful than, for example, looking for periodic patterns. However, since AS routes are frequently deleted in addition to being added, growing networks might not be the best representation.

Table 2.1 illustrates how the same physical system can be represented as different types of networks, often starting with the same initial data.

2.1.3 Network measures

A number of graph theoretic measures are used to characterize the structure networks. In this subsection, we survey the measures that are commonly used for characterizing the graphical properties of networks over

System	*Static*	*Growing*	*Fully dynamic*
Co-auth.	[Newman, 2001c]	[Barabási et al., 2002]	[Moody, 2004]
E-mail	[Ebel et al., 2002]	[Leskovec et al., 2007]	[Kossinets and Watts, 2006][a]
Online	[Mislove et al., 2007]	[Holme et al., 2004]	[Hu and Wang, 2009]

[a] Time-weighted dynamic network.

Table 2.1: EXAMPLES OF PHYSICAL SYSTEMS PAIRED WITH DIFFERENT NETWORK MODELS.

time. We describe these measures to demonstrate that they are non-trivial and diverse, and to motivate the form of randomized sensitivity analysis that we propose in later sections, which is agnostic to the chosen measure.

In addition to classical graph theoretic measures like the average and maximum shortest path lengths between vertices, sociologists have developed a variety of *vertex centrality* measures to assess the importance of individual vertices within the larger network [Wasserman and Faust, 1994], physicists and graph theoreticians have proposed looking at *distributions* of fundamental properties [Newman, 2003, Albert and Barabási, 2002, Erdös and Rényi, 1959], and computer science has seen the wide-scale adoption of ranking techniques like PageRank [Brin and Page, 1998, Langville et al., 2008] and tensor factorization [Sun et al., 2007, Acar et al., 2009] for analysis, and algorithmic questions posed on real networks, such as routing [Kleinberg, 2000].

We focus on three basic categories of network properties that have been widely used to characterize networks that change over time: (1) *connectivity*, in terms of statistics of the distribution of shortest path lengths between all pairs of vertices, (2) the *density* of the network, both local and global, in terms of the relative numbers of vertices and edges, and (3) *spectral* properties of the graph's adjacency matrix or transformations of it.

Connectivity

The *distance* between two vertices u and v in a graph is generally the length of the shortest path between the vertices, *i.e.,* the minimum number of edges that need to be traversed to connect u to v. Let σ_{uv} be the length of the shortest path between vertices u and v, where $\sigma_{uv} = \infty$ if there is no path between u and v (note that σ_{uv} and σ_{vu} are not necessarily equal for directed networks).

Definition 2.1.4. The *connectivity* of a network can be expressed in terms of summary statistics of the pairwise shortest path length distribution, *i.e.,* the distribution of σ_{uv} for all distinct, ordered vertex pairs in a directed network, and for all distinct unordered vertex pairs in an undirected network. In the following definitions, let $k = 1$ for directed networks and $k = 2$ for undirected networks.

$$\bar{l} = \frac{k}{V(V-1)} \sum_{u,v \in V} \sigma_{uv} \qquad \text{(Average path length [West, 2001])}$$

$$\bar{l}^{-1} = \frac{k}{V(V-1)} \sum_{u,v \in V} \frac{1}{\sigma_{uv}} \qquad \text{(Efficiency [Latora and Marchiori, 2001])}$$

$$d_{\max} = \max_{u,v \in V} \sigma_{uv} \qquad \text{(Diameter [West, 2001])}$$

$$d_{90} \approx \arg_i [P(\sigma_{uv} \leq i) = 0.9] \qquad \text{(Effective diameter [Leskovec et al., 2005])}$$

The distribution of all-pairs shortest path lengths between vertices is only well defined for (strongly)

connected graphs, since $\sigma_{uv} = \infty$ when there is no path between u and v. In order to handle disconnected graphs, the average path length distribution is generally only computed over pairs of vertices that are reachable from each other. The *efficiency* measure introduced by Latora and Marchiori [Latora and Marchiori, 2001] is another way to overcome this problem by considering the inverse of the shortest path length, so that when $\sigma_{uv} = \infty$, then $l_{uv}^{-1} = 0$. A third method is to only compute shortest paths between vertices in the largest (strongly) connected component.

The *diameter* of a graph, although unambiguously defined in graph theory [Bollobás, 1998, West, 2001], has a slightly different meaning in other disciplines, particularly physics. For example, in their seminal paper on estimating the 'diameter' of the World Wide Web, Albert *et al.* [Albert et al., 1999] are referring to the average shortest path length and not the maximum shortest path length, as is standard in graph theory. In computer science, Leskovec *et al.* [Leskovec et al., 2005, Leskovec et al., 2007] and subsequently Ahn *et al.* [Ahn et al., 2007] use a smoothed notion of the conventional graph-theoretic definition of diameter – the 90^{th} percentile of the shortest path length distribution – to estimate the length of the 'almost' longest shortest path length in a network. The distinction between these properties is particularly important when considering trends over time, since it is possible to construct examples where, say, the average path length is increasing while the diameter or effective diameter is decreasing.[2]

Density

Irrespective of the structure of a graph, it is sometimes useful to quantify the relative numbers of vertices and edges to get some sense of the 'crowdedness' of the network, either at a global or a local scale. Although density measures are trivial to compute, the ratio of nodes to edges can be particularly informative of structure of the network. For example, a classic paper by Erdös and Rényishowed that the connectivity properties of random graphs undergo an abrupt phase change at certain critical densities that can be analytically computed [Erdös and Rényi, 1959].

Definition 2.1.5. The density of edges in a network $G = (V, E)$ can be quantified in the following ways. In the following definitions, let $N(v)$ be the set of neighbors of vertex v, *i.e.,* the set of vertices that are adjacent to vertex u, and let $k = 1$ for directed networks and $k = 2$ for undirected networks.

$$D = \frac{k|E|}{|V|(|V|-1)} \qquad \text{(Density [Bollobás, 1998, West, 2001])}$$

$$\overline{N} = \frac{1}{|V|}\sum_{v \in V}|N(v)| \qquad \text{(Average degree)}$$

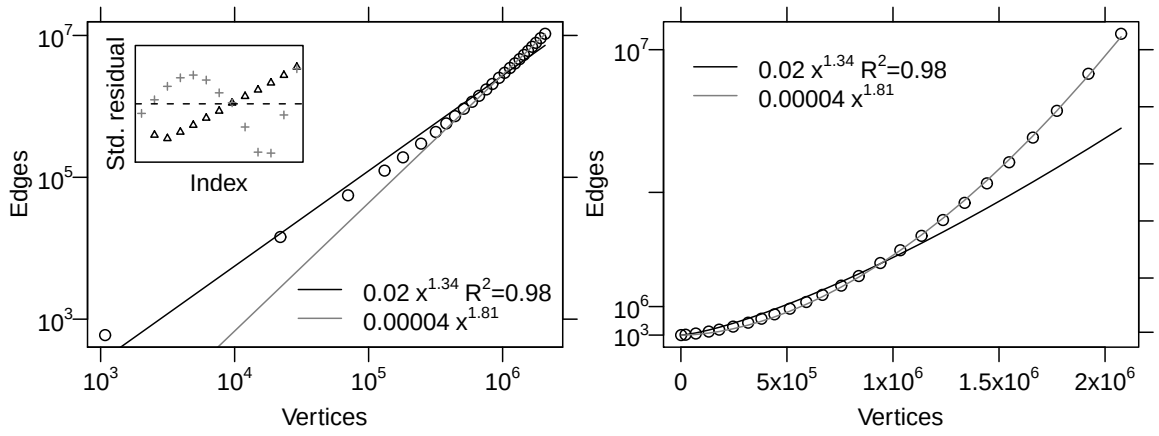
$$= \frac{D}{|V|-1}$$

$$\overline{CC_1} = \frac{1}{|V|}\sum_{v \in V}\frac{k|M(v)|}{|N(v)|(|N(v)|-1)} \qquad \text{(Clust. coefficient [Watts and Strogatz, 1998])}$$

where $M(v)$ is the set of edges between the neighbors of v:

$$M(v) = \{(i, j) : i, j \in N(v) \text{ and } (i, j) \in E\}$$

[2] As a real world example of such a case, see the early stages of the evolution of the *Cyworld* network [Ahn et al., 2007].



(a) Log-log scale

(b) Linear scale

Figure 2.2: DPL plot from a patent citation dataset on doubly logarithmic and linear scales, illustrating the difference between regression assumptions. The black line is a fit of the *Multiplicative* DPL equation, and the gray line is a fit of the *Additive* DPL equation. The inset shows standardized residuals of both fits against a $y = 0$ line, revealing errors that are unlikely to be independent.

and $\overline{CC_1}$ is only computed for vertices with $N(v) > 1$.

The *clustering coefficient* measure proposed by Watts and Strogatz [Watts and Strogatz, 1998] is a measure of the local density of edges around a particular vertex. The clustering coefficient of the whole network is usually expressed as the mean clustering coefficient computed over all vertices. Soffer and Vazquez [Soffer and Vázquez, 2005] show that the clustering coefficient of a vertex, as defined above, is inherently correlated with the degree of the vertex, since high-degree vertices will have more possible edges between them, and a generally lower clustering coefficient. They propose an alternative definition that removes this bias. However, to the best of our knowledge, it has not been widely used.

Finally, a property that has attracted considerable interest in computer science is the *Densification Power Law* (DPL), first described in a pair of seminal papers [Leskovec et al., 2005, Leskovec et al., 2007], examined in a number of subsequent studies [Shi et al., 2007, Menezes et al., 2009, Latapy and Magnien, 2008, Pallis et al., 2009], and even used as the basis of a network sampling algorithm [Leskovec and Faloutsos, 2006].

Definition 2.1.6. An evolving network that obeys a *Densification Power Law* (DPL) contains a number of edges E that, at any point in time, is related to the number of vertices V in a power function of the form:

$$E(t) = k \cdot V(t)^{\alpha} \tag{2.2}$$

where $1 < \alpha < 2$ is called the *densification exponent*.

The DPL is a statement about the relative number of vertices and edges over time. Note that the area of fitting power-law *distributions* to the degrees of vertices measured in a dataset has been well studied [Clauset et al., 2009, Goldstein et al., 2004], but in the case of the DPL, we are not dealing directly with a statistical distribution, but rather a power function of the number of edges to the number of nodes that is invariant over time. A practical issue, however, is that there are at least two ways to fit a power function to empirical

data, each of which requires different assumptions about the underlying relationship, and each of which leads to a different statistical formalization of Definition 2.1.6:

1. *Linear regression on log-transformed data.* This is the regression method that has commonly been used in the literature, including in the original definition [Leskovec et al., 2005]. The DPL function is assumed to have a multiplicative, centered, *i.i.d.*, normally distributed error term ϵ [Seber and Lee, 2003], specifically:
$$E(t) = k \cdot V(t)^{\alpha} \cdot \epsilon$$

2. *Nonlinear regression.* This method assumes a more common additive error term, and requires the use of an algorithm like nonlinear least-squares [Ryan, 2008]:
$$E(t) = k \cdot V(t)^{\alpha} + \epsilon$$

Figure 2.2 illustrates this distinction on a dataset of U.S. patent citations [Hall et al., 2001] (see [Leskovec et al., 2005] for details on preprocessing). The densification exponent of $\alpha = 1.34$ obtained using linear regression is significantly different from the nonlinear regression fit of $\alpha = 1.81$, with the magnitude of the standardized residuals shown in the inset (*i.e.,* differences between the data and fitted curve, centered around the mean difference and divided by the standard deviation in differences). Later in this chapter, we show that each form of DPL exhibits different sensitivities to the same missing data, making it important to qualify the regression technique (and thus statistical noise model) used.

Spectral properties

Spectral graph theory uses the algebraic properties of a graph's adjacency matrix, or transformations of it, to reveal an astonishing amount of information about the structure of the graph [Biggs, 1993, Chung, 1997]. A number of diverse questions can be answered by examining either the set of eigenvalues of the graph's (possibly transformed) adjacency matrix, or specific eigenvectors associated with them. For example, the eigenvector corresponding to the largest eigenvalue is the basis of a number of vertex importance ranking algorithms [Bonacich and Lloyd, 2001, Perra and Fortunato, 2008]. The popular PageRank algorithm can be seen as a variant of this method that uses a transformation of the adjacency matrix [Brin and Page, 1998]. We briefly describe some interesting spectral graph properties, focusing on eigenvalues as a global characterization of the graph rather than on eigenvectors as characterizations of individuals vertices.

Definition 2.1.7. The *spectrum* of a matrix is the set of its eigenvalues. The spectral properties of a graph $G = (V, E)$ depend upon which matrix representation is chosen for it. Starting with the basic adjacency matrix of a graph, the *combinatorial Laplacian matrix* is defined as follows, where $N(v)$ is the degree of vertex $v \in V$:
$$L_{uv} = \begin{cases} N(v) & : u = v \\ -1 & : (u, v) \in E \\ 0 & : \text{otherwise} \end{cases}$$

We denote $\lambda_1 \leq ... \leq \lambda_V$ as the ordered set of eigenvalues of an adjacency matrix A, where V is the number of vertices in the graph and λ_V is the *principal eigenvalue*. Similarly, the spectrum of the Laplacian matrix is denoted $\mu_1 \leq ... \leq \mu_V$.

The following are some properties of the spectra of each matrix representation of a graph.

1. *(Adjacency)* The diameter of the graph $d_{\max}$ is less than the number of distinct eigenvalues [West, 2001].

2. *(Adjacency)* For processes spreading on graphs, such as viruses in computer networks, the *epidemic threshold* is a critical rate of infection beyond which a virus can become an epidemic. The epidemic threshold can be estimated by the quantity $1/\lambda_V$ [Wang et al., 2003].

3. *(Adjacency)* In a scale-free graph, the maximum degree grows as $\sqrt{V}$, so the principal eigenvalue can be expected to grow as $\lambda_V \sim V^{1/4}$ [Farkas et al., 2001].

4. *(Laplacian)* The number of zero eigenvalues is equal to the number of connected components.

5. *(Laplacian)* The largest eigenvalue is bounded by twice the maximum degree [Almendral and Díaz-Guilera, 2007]:

$$\mu_V \leq 2N_{\max}$$

6. *(Laplacian)* The diameter of an undirected graph $d_{\max}$ is bounded by the following expression involving the largest and second-smallest eigenvalue μ_2 [Chung et al., 1994]:

$$d_{\max} \leq \left\lfloor \frac{\cosh^{-1}(V-1)}{\cosh^{-1}(\frac{\mu_V+\mu_2}{\mu_V-\mu_2})} \right\rfloor + 1,$$

Spectral properties have rarely been use to characterize the time evolution of networks, perhaps because interpreting temporal trends in eigenvalues is less intuitive than interpreting trends in classical graph theoretic properties like density and diameter. Plotting the spectrum of a graph can be used as an effective tool to summarize the structure of the graph [Banerjee and Jost, 2009], but robust and intuitive methods for tracking this summary over time are an open problem.

2.2 Dynamic properties of real networks

In the previous section, we outlined network data classes, aggregation methods, and graph measures used to characterize network structure. Using this framework, we now survey the literature to summarize commonly reported temporal properties of real dynamic network datasets. We also aim to classify the experimental methodology of each study into the conceptual framework described in the previous section, in an attempt to characterize what the *de-facto* experimental procedure is. Reported empirical results are categorized by graph-theoretic properties and the type of network dataset being studied, and then presented in increasing chronological order of year of publication. Section 2.2.1 describes the literature in terms of global and local density measures in the graph, whereas Section 2.2.2 describes distance measurements in terms of the distribution of shortest paths over time. Due to the ordering of this section, the networks of each study are described in most detail in Section 2.2.1.

2.2.1 Global and local density

We first examine three measures related to the global and local density of a network: the conventional definition of density D (the ratio of the number of edges to the maximum possible number of edges), the average degree $\overline{N}$, and the average clustering coefficient $\overline{CC_1}$. Of these, $\overline{CC_1}$ measures the local density of edges centered at each vertex in the network, on average, as specified in Definition 2.1.5.

Bibliographic networks

Barabási *et al.* [Barabási et al., 2002] analyzed the time evolution of two co-authorship datasets consisting of publications in Neuroscience and Mathematics. They found that the average degree $\overline{N}$ in both datasets are monotonically increasing over time, although the rate of increase is faster in Neuroscience than in Mathematics. They speculate that this is because of the differences in collaboration culture between the fields. The average clustering coefficient is also decreasing in both datasets, although this could possibly be explained by the degree-correlation bias in the $\overline{CC}$ measure [Soffer and Vázquez, 2005].

Elmacioglu and Lee [Elmacioglu and Lee, 2005] presented an analysis of co-authorship in the database research community by aggregating co-authorship information from 100 "hand-picked" publication venues listed in the DBLP database [Ley, 2002]. We obtained two trends related to density from their analysis. The first is the average degree over time considering the DBLP database as a dynamic network, expressed as the average number of collaborators per author each year. This trend is shown to be increasing, and is a lower bound on the trend in a growing version of the same dataset. The second trend in the paper is a slow, linear increase in the clustering coefficient over time computed in the growing (cumulative) version of the dataset. The apparent disagreement with the trend found by Barabási *et al.* in other co-authorship datasets could possibly be explained by the fact that Elmacioglu and Lee only compute the clustering coefficient for vertices in the largest connected component of the network, whereas Barabási *et al.* consider all vertices in the network. At the end of the observation period, less than 60% of all vertices in Elmacioglu and Lee's dataset are included in the giant component.

Menezes *et al.* [Menezes et al., 2009] analyzed the structure and evolution of research collaborations in computer science from co-authorship data. They manually selected academic institutions from three geographical regions – 16 in the United States and Canada (Ca-US), 6 in Europe (Fr-Sw-UK), and 8 in Brazil (Br) – and extracted names of faculty from departmental homepages, and co-authorship data for those faculty from DBLP. In the Brazil and Europe networks, the clustering coefficient appears to be decreasing at a uniform rate. However, the North American network appears to be largely stable, subject to regular fluctuations. This is in contrast to the fraction of vertices in the giant component, where both Brazil and Europe exhibit sharp increases.

Online social networks

Holme *et al.* [Holme et al., 2004] obtained an almost exhaustive history of user interactions in *Pussokram*, an online dating social network popular in Sweden. This network is somewhat unusual because it contains full temporal information about multiple modes of user interactions. Users can, for example, list each other as 'friends', or write private or public messages to each other, or 'flirt' with one another, and each type of interaction generates a different network. We report the authors' results on the aggregate network of all

interactions and on friendship links alone, since the other trends are very similar. All interaction types yield networks that exhibit an initially increasing average degree, but in all cases, the trend appears to be quickly converging to a constant. In the case of friendship links, the average degree is approximately constant for more than 400 out of 500 days in the observation period. The authors note that this appears to agree with sociological constraints, such as empirical findings that maximum social network sizes in humans appear to be bounded [Hill and Dunbar, 2003].

Local density in the *Pussokram* network also exhibits some interesting trends. The authors compute both directed and undirected clustering coefficient in the network over time, and show that the network of all interactions exhibits a decreasing clustering coefficient, similar to earlier findings in bibliographic networks by Barabási *et al.* [Barabási et al., 2002]. The network consisting only of friendship links, however, displays a non-monotonic trend, and it is hard to draw any conclusions about long-term behavior. The authors note that since *Pussokram* is primarily a dating site, conventional social norms for introducing one's friends to each other, and thus increasing local density, might not apply.

Kumar *et al.* [Kumar et al., 2006] obtained complete temporal information on the evolution of two online social networks - *Yahoo 360* and *Flickr*, both of which are treated as growing networks. A key finding, which can perhaps be generalized to other online social networks where complete temporal information is available, is that there appear to be multiple regimes of growth. The authors note that these stages correspond to an "initial euphoria" peak as early adopters sign up for the service, a subsequent decline of the initial enthusiasm, followed by steady "organic growth". Both networks exhibit these phenomena in terms of the average degree of nodes, with the long-term trend appearing to be an increasing one.

Beyene *et al.* [Beyene et al., 2008] examine the structure of a *trust network*, built from binary feedback ratings on the online auction site eBay. These ratings are assigned to users by other users on the completion of a transaction, and can be either positive or negative, although they are generally found to be positive. Beyene *et al.* show that the average degree in this network increases almost linearly with time, from 1999 to 2005. Note that the data was collected by crawling user profiles on the eBay site, which brings up issues of incomplete data, and consequently, the missing past.

Hu *et al.* [Hu and Wang, 2009] analyze the temporal evolution of *Wealink*, an online social network popular in China, as a dynamic network. Perhaps the most interesting feature of their dataset is the S-shaped trend, resembling a logistic function, of both the number of vertices and edges over time. Since the dataset is an exhaustive history of the evolution of the network, the authors speculate that the sharp increase in the number of vertices and edges corresponds to a sudden burst of popularity, when the membership of the network grew from under 10,000 vertices to over 200,000 in less than 5 months out of a 27 month history, similar to earlier findings by Kumar *et al.* [Kumar et al., 2006]. Both the global density D and the average clustering coefficient $\overline{CC_1}$ appear to decrease towards a constant once the growth of the network has stabilized. A fit to the logistic function suggests that the number of vertices has an asymptotic limit of $V \sim 224,000$ and $E \sim 272,100$.

Subsets of the World Wide Web

Buriol *et al.* [Buriol et al., 2006] analyzed the evolution of the structure of hyperlinks between articles in *Wikipedia*, an online, publicly-editable encyclopedia. Using snapshots of the link structure that model Wikipedia as a dynamic network, they find that the average out-degree $\overline{N_{out}}$ is increasing at a constant rate

of approximately one new out-link every 100 days. The average clustering coefficient remains approximately constant in the last year of network evolution, after periods of non-monotonicity. This is somewhat surprising, since it implies that the local density around vertices remains constant even though the average number of neighbors increases.

Shi *et al.* [Shi et al., 2007] analyzed four temporal snapshots of a partial crawl of blogspace released as part of the TREC Blog-Track 2006 dataset. They report that that the average degree $\overline{N}$ of the network increases from 1.5 to 2.657 over a period of 40 days. Similarly, $\overline{CC_1}$ increases monotonically from 0.034 to 0.052 over the same period.

Latapy and Magnien [Latapy and Magnien, 2008] analyze a crawl of the `.uk` WWW domain, where each hyperlink is labeled with the time that it is discovered. Note that the motivation for the study was not to study the evolution of the network, but to determine how large a network sample should be in order for network properties to stabilize. In order to achieve this, the authors add edges sequentially to a network in increasing order of the timestamp of the edge. Thus, their methodology essentially builds a growing network, allowing us to compare their results, at least qualitatively, to studies that explicitly analyze network properties over time. Although the timestamp on the hyperlinks in the WWW dataset correspond to the link's discovery time and not the link's creation time, we include the datasets in this survey since the same methodology has been used in other papers [Leskovec et al., 2007]. The average degree grows quickly as a function of the number of nodes added to the network, but the density appears to decrease smoothly and the clustering coefficient is extremely non-monotonic. The Web dataset also seems to show sharp discontinuities in time, perhaps implying problems in the underlying crawling process. Thus, the trends in this particular dataset should not be given too much importance.

Internet Routers and Autonomous Systems

Dhamdhere and Dovrolis [Dhamdhere and Dovrolis, 2008] analyze the structure and evolution of a specific type of link between Autonomous Systems, namely 'customer-provider' (CP) or 'paid transit' links instead of all links. They perform extensive filtering and smoothing on the dataset, and show that the limited nature of AS datasets nonetheless allows reasonable estimation of the topology of CP links, but not all links. The average degree $\overline{N}$ over CP links is shown to be steadily increasing over time.

Dynamic Interaction Networks

Kossinets and Watts [Kossinets and Watts, 2006] analyzed a dynamic network of e-mail communications between university students, staff, and faculty. Since e-mail constitutes a dynamic interaction network, they used decay windows of various lengths in order to smooth network structure. Specifically, each edge between vertices i and j is weighted at time t according to the following function[3]:

$$w_{ij} = \frac{\sqrt{m_{ij} m_{ji}}}{\tau}$$

where τ is the length of a chosen smoothing time window, and m_{ij} is the number of messages sent from i to j in the period $(t - \tau, t]$. For vertices i and j, if $w_{ij} > 0$ at time t, then the edge (i, j) exists in the network at that time. This introduces a form of edge decay over time, since e-mails allow the observation of

[3]Found in the supporting online material for [Kossinets and Watts, 2006], available at `www.sciencemag.org`.

the creation of links, but not of their dissolution. It also implements the intuitive idea that an e-mail should not constitute a social tie in a network indefinitely.

Using this edge decay scheme and window sizes of $\tau = 30, 60, 90$ days, the authors find that the average degree $\overline{N}$ is approximately constant during the semester, decreasing sharply during the semester break and in summer. Similarly, the average clustering coefficient $\overline{CC}$ appears to be largely in equilibrium during the semesters, its trend only increasing slightly during summer. The drops in average degree during summer, for example, can be explained by students not being on campus. Although the authors were only able to obtain data for one academic year, their results raise a number of interesting points: namely, that a single, global smoothing parameter appears to show that the network is largely in equilibrium during the semesters. Whether this phenomenon could be observed in other networks using a similar smoothing mechanism is an open question.

Pallis *et al.* [Pallis et al., 2009] analyze the structure of dynamic vehicular ad-hoc networks, *i.e.,* temporary wireless networks created when specially-equipped vehicles on open roads are within transmission range of each other. Their study is notable because it does not use actual interaction data, but rather the results of a large-scale realistic traffic simulation [Raney et al., 2002, Naumov et al., 2006]. The authors simulate traffic in the center of the city of Zurich, Switzerland for 3 hours in the morning rush period, and do not use any form of edge decay or sliding window, as in Kossinets and Watts [Kossinets and Watts, 2006]. Since the time evolution of the network reflects instantaneous traffic patterns, it is not surprising that the average degree increases gradually as traffic starts to build, and then falls off. The clustering coefficient, on the other hand, remains constant throughout the simulation period, making this an example of a graph where local and global density appear to uncorrelated.

Pallis *et al.* [Pallis et al., 2009] also illustrate a somewhat unusual usage of the DPL relationship in their analysis of communication links in dynamic vehicular ad-hoc networks. We have mentioned that the authors do not use any form of edge decay or smoothing, resulting in an instantaneous picture of communication links at each timestep in the evolution of the network. The authors fit the Multiplicative DPL to the model and find a densification exponent of $\alpha = 1.77$, although the interpretation of this value is difficult.

Year	Reference	Network	Type	Category	$\overline{N}$	D	$\overline{CC}$
2002	[Barabási et al., 2002]	Mathematics	Biblio.	growing	increasing	-	decreasing
		Neuroscience	Biblio.	growing	increasing	-	decreasing
2004	[Holme et al., 2004]	Pussokram: *All*	Online	growing	increasing[a]	-	decreasing
		Pussokram: *Friends*	Online	growing	constant	-	[b]
2004	[Park et al., 2004]	RouteViews	AS	dynamic	increasing[b]	-	increasing
		Extended	AS	dynamic	-	-	-
2005	[Elmacioglu and Lee, 2005]	DBLP-DB	Biblio.	growing	increasing[c]	-	increasing[b]
2006	[Kossinets and Watts, 2006]	University	Email	interaction	periodic[d]	-	periodic[d]
2006	[Buriol et al., 2006]	Wikipedia	WWW	dynamic	increasing	-	constant[b]
2006	[Kumar et al., 2006]	Flickr	Online	growing	increasing[b]	-	-
		Yahoo 360	Online	growing	increasing[b]	-	-
2007	[Shi et al., 2007]	TREC	WWW	dynamic	increasing	-	increasing
2008	[Dhamdhere and Dovrolis, 2008]	AS-CP	AS	dynamic	increasing[b]	-	-
2008	[Latapy and Magnien, 2008]	INET	Router	growing	increasing	[b]	increasing[a,b]
		eDonkey	P2P	growing	increasing	decreasing[a,b]	decreasing[b]
		Metrosec	Router	growing	constant[b]	decreasing[a,b]	decreasing[a,b]
2008	[Beyene et al., 2008]	eBay	Online	growing	increasing	-	-
2009	[Menezes et al., 2009]	Brazil	Biblio.	growing	-	-	decreasing
		Ca-US	Biblio.	growing	-	-	decreasing
		Fr-Sw-UK	Biblio.	growing	-	-	constant[b]
2009	[Hu and Wang, 2009]	Wealink	Online	dynamic	-	constant[b]	constant[b]
2009	[Pallis et al., 2009]	Vehicle traffic	Ad-hoc	interaction	[b]	-	constant

[a] Trend suggests convergence to asymptote.

[b] Trend exhibits non-monotonic behavior.

[c] Trend measured using fully dynamic network aggregation.

[d] Uses smoothing, sliding windows, or edge decay with fully dynamic network aggregation.

Table 2.2: OVERVIEW OF LOCAL AND GLOBAL DENSITY MEASUREMENTS. NOTE THAT THESE ARE QUALITATIVE ASSESSMENTS BASED ON GRAPHICAL DATA PRESENTED IN EACH PAPER.

Year	Reference	Network	Type	Category	$\bar{l}$	$d_{\max}$	d_{90}
2002	[Barabási et al., 2002]	Mathematics	Biblio.	growing	decreasing[a]	-	-
		Neuroscience	Biblio.	growing	decreasing[a]	-	-
2003	[Nascimento et al., 2003]	SIGMOD	Biblio.	growing	[b]	constant[b]	-
2004	[Holme et al., 2004]	Pussokram: *All*	Online	growing	decreasing[a]	-	-
		Pussokram: *Friends*	Online	growing	increasing[b]	-	-
2004	[Park et al., 2004]	RouteViews	AS	dynamic	decreasing	-	-
		Extended	AS	dynamic	[b]	-	-
2005	[Elmacioglu and Lee, 2005]	DBLP-DB	Biblio.	growing	constant[b]	-	-
2006	[Kumar et al., 2006]	Flickr	Online	growing	constant[b]	-	increasing[b]
		Yahoo 360	Online	growing	decreasing[b]	-	decreasing[b]
2006	[Kossinets and Watts, 2006]	University	Email	interaction	periodic[d]	-	-
2007	[Ahn et al., 2007]	Cyworld	Online	dynamic	decreasing[b]	-	decreasing[b]
2008	[Dhamdhere and Dovrolis, 2008]	AS-CP	AS	dynamic	constant[b]	-	-
2008	[Latapy and Magnien, 2008]	INET	Router	growing	decreasing[a]	decreasing[a]	-
		eDonkey	P2P	growing	decreasing[a]	decreasing	-
		Metrosec	Router	growing	constant	constant	-
2009	[Menezes et al., 2009]	Brazil	Biblio.	growing	[b]	-	-
		Ca-US	Biblio.	growing	decreasing	-	-
		Fr-Sw-UK	Biblio.	growing	[b]	-	-
2009	[Hu and Wang, 2009]	Wealink	Online	dynamic	decreasing[a b]	constant[b]	-
2009	[Pallis et al., 2009]	Vehicle traffic	Ad-hoc	interaction	noisy	-	-

[a] Trend suggests convergence to asymptote.

[b] Trend exhibits non-monotonic behavior.

[c] Trend measured using fully dynamic network aggregation.

[d] Uses smoothing, sliding windows, or edge decay with fully dynamic network aggregation.

Table 2.3: OVERVIEW OF AVERAGE AND MAXIMUM DISTANCE MEASUREMENTS. NOTE THAT THESE ARE QUALITATIVE ASSESSMENTS BASED ON GRAPHICAL DATA PRESENTED IN EACH PAPER.

Year	Reference	Network	Category	Type	Time span	Data	Miss. past
2002	[Barabási et al., 2002]	Mathematics	Biblio.	growing	1991-1998	8	yes
		Neuroscience	Biblio.	growing	1991-1998	8	yes
2003	[Nascimento et al., 2003]	SIGMOD	Biblio.	growing	1975-2002	28	yes
2004	[Holme et al., 2004]	Pussokram	Online	growing	512 days	large	yes
2004	[Park et al., 2004]	RouteViews	AS	dynamic	1997-2002	large	yes
		Extended	AS	dynamic	∼7 weeks	9	yes
2005	[Elmacioglu and Lee, 2005]	DBLP-DB	Biblio.	growing	1968-2003	36	yes
2005	[Leskovec et al., 2005]	arXiv	Biblio.	growing	1993-2003	124	yes
		Patents	Citation	growing	1975-1999	25	yes
		AS	AS	dynamic	1997-2000	735	yes
		Affiliation	Biblio.	growing	1992-2002	10	yes
2006	[Kossinets and Watts, 2006]	University	Email	interaction	∼1 year	large	yes
2006	[Buriol et al., 2006]	Wikipedia	Web	dynamic	2002-2006	17	yes
2006	[Kumar et al., 2006]	Flickr	Online	growing	100 weeks	100	no
		Yahoo 360	Online	growing	40 weeks	40	no
2007	[Ahn et al., 2007]	Cyworld	Online	dynamic	2002-2006	8	yes
2007	[Shi et al., 2007]	TREC	Blogs	growing	40 days	4	yes
2007	[Leskovec et al., 2007]	Email	Email	growing	18 months	18	yes
		IMDB Actors-Movies	General	growing	1890-2004	114	no
2008	[Latapy and Magnien, 2008]	INET	Router	growing	16 months	large	yes
		eDonkey	P2P	growing	47 hours	large	no
		Metrosec	Router	growing	8 days	large	yes
2008	[Dhamdhere and Dovrolis, 2008]	AS-CP	AS	dynamic	1998-2007	40	yes
2008	[Huang et al., 2008]	CiteSeer	Biblio.	growing	1980-2005	25	yes
2008	[Beyene et al., 2008]	eBay	Online	growing	1999-2005	7	yes
2009	[Menezes et al., 2009]	Brazil	Biblio.	growing	1994-2006	13	yes
		Ca-US	Biblio.	growing	1994-2006	13	yes
		Fr-Sw-UK	Biblio.	growing	1994-2006	13	yes
2009	[Hu and Wang, 2009]	Wealink	Online	dynamic	2005-2007	27	no
2009	[Dong et al., 2009]	China	Airport	dynamic	1983-2006	3	(?)
2009	[Pallis et al., 2009]	Vehicle traffic	Ad-hoc	interaction	3 hours	large	no

Table 2.4: AN OVERVIEW OF THE CHARACTERISTICS OF EVOLVING NETWORK DATASETS.

2.2.2 Connectivity

Bibliographic networks

Barabási *et al.* [Barabási et al., 2002] were among the first to find a decreasing average shortest path length $\bar{l}$ in a growing network, which does not agree with network growth models like Preferential Attachment [Newman, 2001b]. For both the Mathematics and Neuroscience datasets that they examine, $\bar{l}$ is decreasing and apparently converging to an asymptote. The authors note that a longer observation interval might indicate that $\bar{l}$ approaches a stationary value, but the relative novelty of co-authorship datasets at the time of publication resulted in just 8 data points, with each representing an additional year of cumulative co-authorships. Note that although the authors refer to 'diameter', the only quantity studied is the average shortest path length $\bar{l}$.

Nascimento *et al.* [Nascimento et al., 2003] analyzed the co-authorship graph of the SIGMOD conference as a growing network. The average path length in the largest connected component in the network varies considerably over the years, but eventually settles down into what appears to be a decreasing trend. We have not listed any trend in Table 2.3 because of the extremely short time period that any trend is visible at all. The authors note that computing path lengths within the largest component only has its caveats: until 1980, only 16 authors were in the largest connected component, out of the 1,683 that eventually appear. The trend is $\bar{l}$ is therefore predictably noisy. The authors also report that $d_{\max}$ in the network has been constant at a value of 15 between 1996 and 2002.

Elmacioglu and Lee [Elmacioglu and Lee, 2005] analyzed the growing network of publications in the database research community, constructed from a set of manually selected publication venues reported to the DBLP database. They report the value of the average shortest path length $\bar{l}$ over time, computed over all reachable pairs of vertices in the network. After an initial increasing burst, the trend in $\bar{l}$ appears to stabilize to a constant value around 6 from approximately 1989 to 2003. While it is certainly possible that database research was particularly energized between 1973 and 1983, resulting in an infusion of new authors and sharply increasing $\bar{l}$, it is also possible that this spike is an artifact caused by missing past issues, or known limitations in the indexing process of DBLP for early data [Ley, 2002]. However, the fact that $\bar{l}$ appears to stabilize for a number of years would suggest that the network has reached an equilibrium.

Menezes *et al.* [Menezes et al., 2009] analyzed three co-authorship networks constructed from faculty publications at manually selected universities in North American, Europe, and Brazil. The average shortest path length $\bar{l}$ in the North American network displays a relatively uniform decreasing trend, but both the Brazilian and European networks exhibit sharp increases by doubling between 1998 and 2001. Following these sharp increases, the Brazilian network appears to settle into a decreasing trend, whereas the European network continues increasing. It should be noted that in the Brazilian research network, the doubling of $\bar{l}$ is correlated with a sharp increase in the fraction of nodes in the giant component.

Online social networks

In their analysis of the *Pussokram* online dating social network, Holme *et al.* [Holme et al., 2004] find different trends for networks built from different types of interactions. Note that although the authors have complete history of the network, registered users of a different service had their accounts "automatically transferred to pussokram.com" on its inception [Holme et al., 2004], which implies that this dataset also

has a version of the missing past issue. The authors report different trends in the average shortest path length $\bar{l}$, computed only within the largest giant component as opposed to between all pairs of reachable nodes, depending on the type of interaction used to build the network. The aggregate network built from all possible user interactions, for example, shows a decreasing trend in $\bar{l}$, appearing to converge to a constant, but the network built only from friendship links displays an *increasing* trend after a very brief period of initial decrease. One possibility is that each type of user interaction on the social network is governed by a different process, which results in different trends in each network, but it is also possible that the differences are caused by disparities in the amount of data on each type of interaction.

Kumar *et al.* [Kumar et al., 2006] analyzed social networks of the *Yahoo 360* and *Flickr* services as growing networks. Of these, the *Flickr* network is somewhat unusual in terms of both the $\bar{l}$ and d_{90} measures. In the "organic growth" phase, it is difficult to discern a long-term trend in either measure from graphical data. Although the authors state that both measures decrease over time, there appears to be a slight, monotonic *increase* in both measures for approximately the last 30 weeks out of 100. The *Yahoo 360* network also exhibits noisy trends, although the long-term trend in both $\bar{l}$ and d_{90} appear to be decreasing. Both these networks are exception because complete temporal information is available for every network-altering activity. They also illustrate the problems associated with characterizing the properties of real networks as simple, monotonic trends.

Ahn *et al.* [Ahn et al., 2007] present an analysis of the evolution of friendship links in *Cyworld*, an online social network popular in South Korea. Although they claim to have obtained the complete topology of the network directly from the service provider, the amount of temporal information is limited to about 42% of the data. Thus, the missing past issue is a consideration in the temporal analysis of this dataset. The authors find that the average shortest path length $\bar{l}$ increases almost linearly for the first three and a half years of data, but then gradually starts to drop. The effective diameter d_{90}, on the other hand, stays approximately constant during the early period, before dropping off sharply and almost converging to the $\bar{l}$ trend. Note that the drop in both $\bar{l}$ and d_{90}, and particularly the reversal of the trend in $\bar{l}$, coincides with a sharp increase in the number of nodes in the network, which could indicate a change in the evolution dynamics of the network caused by perhaps some sort of external event or marketing effort. In either case, it reinforces the need to consider the underlying processes reflected in the dataset.

Hu *et al.* [Hu and Wang, 2009] note that a complete temporal history of the *Wealink* online social network exhibits non-monotonic behavior in both the average shortest path length $\bar{l}$ as well as the diameter d, *i.e.,* the maximum shortest path length. However, the network underwent a rapid and extremely pronounced burst of growth, after which network properties seem to stabilize. The long-term trend in $\bar{l}$ appears to be slowly decreasing, whereas the diameter d stays constant. As Leskovec *et al.* [Leskovec et al., 2005] point out, the diameter is extremely sensitive to outlier structures, so the effective diameter d_{90} or 'almost' longest path might follow a different trend.

Internet Routers and Autonomous Systems

In their analysis of the evolution of CP links in Autonomous Systems, Dhamdhere and Dovrolis [Dhamdhere and Dovrolis, 2008] show that the average shortest path length $\bar{l}$ has remained approximately constant over the last 9 years.

Dhamdhere and Dovrolis [Dhamdhere and Dovrolis, 2008] analyze the structure and evolution of a specific

type of link between Autonomous Systems, namely 'customer-provider' (CP) or 'paid transit' links instead of all links. They perform extensive filtering and smoothing on the dataset, and show that the limited nature of AS datasets nonetheless allows reasonable estimation of the topology of CP links, but not all links. The average degree $\overline{N}$ over CP links is shown to be steadily increasing over time.

Dynamic Interaction Networks

Kossinets and Watts [Kossinets and Watts, 2006] analyzed a university's e-mail records for one academic year as a smoothed dynamic interaction network, which we have previously described in Section 2.2.1. The authors report two interesting findings. The first is that with smoothing windows of 60 and 90 days, the average shortest path length $\bar{l}$ stays constant, except after the onset of summer, when it rises. This implies that the network is in equilibrium until the summer, when presumably a large fraction of the student population leaves campus and uses e-mail less frequently. The second finding of interest is that for the shortest smoothing window of 30 days, intended to capture fast-changing dynamics, there is a strong correlation between the fraction of vertices in the largest component and the average shortest path length $\bar{l}$ with the largest component. The first phenomenon can again be explained by its occurrence during semester breaks, but the correlation between $\bar{l}$ and the size of the giant component raises the question of the underlying process responsible for the increase or decrease in $\bar{l}$ in other studies that compute $\bar{l}$ within the largest component only.

2.2.3 Summary

In conclusion, empirical evidence seems to suggest the following common trends in real networks, as measured using the growing network methodology:

1. The average shortest path length is decreasing over time, often appearing to converge to a fixed value.
2. The effective diameter decreases over time.
3. The average degree grows over time, apparently without a bound.

In the next two sections, we investigate how sensitive these trends might be to various kinds of sampling error. The next section deals with citation-type networks, and the subsequent section with interaction-type networks.

2.3 Sensitivity of measured trends in citation networks

Recall that citation networks grow over time with the addition of nodes and edges, and that each individual edge can only appear once in the timestream, to represent true growth in the underlying system. When the growing network method is used to aggregate observations of a citation network, an implicit assumption is that every aggregated observation is representative of the underlying system at that point in time, with respect to a measure M, *i.e.,* that the magnitude of change in the underlying system from time t_1 to time t_2 is proportionally represented in a change in measure M from t_1 to t_2, and vice versa.

There is, however, a condition under which this assumption does not hold: when the dataset does not contain a full temporal history of the underlying process. This is called a *missing past*, and a version of

it has been briefly described before in [Barabási et al., 2002] and [Leskovec et al., 2005]. Assume that a physical system is continuously in a state of flux, but that the observations in a citation network dataset of it necessarily start at an arbitrary time $t_0 > 0$. Furthermore, assume that due to limitations in the observation process, the partial picture we have of the underlying system at time t_0, with respect to measure M, is incomplete. Since individual edges are only observed once in a citation network's lifetime, we will never discover the existence of edges that occurred before time t_0.

However, the same is generally not true for discovering vertices that existed before time t_0. When a new vertex joins the network and links to a vertex that existed before t_0, the older vertex is 're-discovered' and can incorrectly be presumed to be a new vertex. We call this phenomenon *vertex re-dicovery*, and it can sometimes be controlled with additional metadata. For example, Leskovec *et al.* [Leskovec et al., 2007] use a patent citation dataset where the time that each node was created is explicitly recorded, eliminating any vertices for which they do not have a creation time. However, when this information is not available, each *apparently* new vertex can affect a measure M by a significant amount, increasing the measurement error between the true and observed networks. In some cases, as we will show, this error can progressively alter trends in some measures over time, suggesting change in the underlying network in a manner that is not actually happening. This is the first source of error in dynamic network measurements that we analyze.

In addition to vertex re-discovery affecting graph measurements, a second source of error is simply the structure of the missing past graph. Even assuming that the measurement process is perfect and accurately records all vertex and edge additions in the observation stream, the unobserved temporal history of the network contains a missing past graph of unknown structure. The true change in the underlying system is determined by the structure of this missing past graph in relation to the actual observations. Unfortunately, this missing data graph looks like is difficult to determine, in general, for real datasets.

2.3.1 Assessing sensitivity

The central question of this section is whether the errors introduced by the processes just mentioned are significant enough to be of concern. Given a dataset and a measure M, the difficulty of assessing the sensitivity of a temporal trend in M is that we would require the structure of the missing past network in order to do so. However, a number of stochastic models of growing networks have already been developed that allow us to generate presumably realistic, continuously growing, citation networks [Chakrabarti et al., 2010]. This allows us to conduct a systematic study of the effects of various amounts of missing past on network measurements by using simulations of these models to generate data. The diversity of network growth models in the literature means that we can systematically study a reasonably large class of dynamical systems.

We use a simulation setup where a network growth model is used to generate a synthetic dynamic network dataset with known properties. We designate this as the ground truth network. By censoring the initial portion of the synthetic network timeline, we can simulate both growth and vertex re-discovery without modifying the network growth model. The independent parameter of the simulation (beyond the parameters of individual network growth models) is the amount of data to censor, which has a well-defined meaning, and is the primary phenomenon we are interested in. Ideally, the missing past effect would be small and the graph properties we study would converge to their true values quickly, or at least yield qualitatively similar trends over time.

There have been two prior attempts, to the best of our knowledge, to investigate the impact of missing past data on network trends over time when using the growing network aggregation method. This is in contrast to the many studies described in Section 2.2 that simply use the growing network method under the same conditions. In 2002, Barabási *et al.* described a supplementary experiment assessing the effect of missing past on the average shortest path length over time, in their seminal paper on co-authorship networks [Barabási et al., 2002]. Our simulation setup in this section is very similar to theirs, but we systematically study several citation network generating processes, graph measures, and amounts of missing past data.

The second experiment was performed in 2007 by Leskovec *et al.*, who use a subtly different definition than Barabási *et al.* of what constitutes the missing past. Namely, they only consider a limited form of the missing past in citation networks: "citations to [vertices] that predate our earliest recorded time" [Leskovec et al., 2007, pg.17]. Where Barabási *et al.* analyzed the effect of missing past using synthetic simulation data, Leskovec *et al.* use a real dataset, known to have a missing past with unknown structure, in order to determine what the impact of the missing past would be. They used the following experiment to validate that the trend they observed in the *effective diameter* graph measure was not an artifact of the observation process (note that $t = 0$ corresponds to the start of observations in the quotation below, not to the start of the process as in our notation):

> We pick some positive time $t_0 > 0$ and determine what the diameter would look like as a function of time if this were the beginning of our data. We then put back in the nodes and edges from before time t_0 and study how much the diameters change. If this change is small – or at least if it does not affect the qualitative conclusions – then it provides evidence that the missing past is not influencing the overall result [Leskovec et al., 2007].

We can now illustrate how simulations with synthetic data can shed light on the efficacy of the test described above. A well-known network generation model is the *preferential attachment*, which we describe in more detail subsequently in Section 2.3.2, and which is known to generate graphs with a slowly growing diameter (see Section 2.1.3 for definition). Starting with an initial seed graph of a single vertex, we use the preferential attachment model to generate a random 'ground truth' growing citation network. After an arbitrary number of timesteps t_0, we simulate the start of the observation process. Specifically, the entire structure of the network prior to t_0 is considered to be the missing past and censored, but additions to its structure are observed and aggregated into a growing graph. We can then compare the value of some measure M, here the effective diameter, on the full ground truth network and the truncated, mising past network. This allows us to measure the impact of different amounts of missing past.

Figure 2.3 shows the ground truth for the d_{90} (effective diameter) measure over time for the true dataset and the measured, missing past network. In this instance, the effect of the missing past is extremely significant – where the ground truth network shows a slowly increasing diameter, the missing past network would suggest that the diameter is rapidly decreasing over time. The inference we make about the change in the structure of the underlying process is therefore an artifact of the structure of the missing past network and vertex re-discovery. This essentially reproduces the results of the experiment conducted by Barabási *et al.* , using effective diameter as the measure of interest instead of average shortest path length [Barabási et al., 2002].

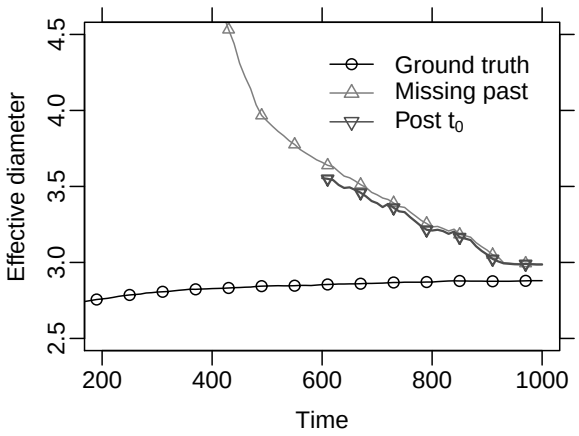


Figure 2.3: The effective diameter over time in a synthetic preferential attachment network, the observed trend after an initial portion is censored to simulate the missing past, and the trend produced by a 'Post-t_0' validation experiment

We then conducted the experiment described by Leskovec *et al.* above on the truncated missing past network, treating it as the observed dataset, and knowing the ground truth in advance. We chose an arbitrary time greater than t_0 and essentially repeated the procedure of truncating the network before that time. The 'Post-t_0' validation experiment used by Leskovec *et al.* intuitively states that if the trend in the missing past network and the 'Post-t_0' are comparable, then the trend in the missing past network and the ground truth are also comparable. However, we find that the trend resulting from the validation experiment is essentially identical to the trend in the missing past network, in agreement with similar observations on real datasets in [Leskovec et al., 2007]. However, both trends are nonetheless artifacts here of the missing past, given that the ground truth has a slowly increasing d_{90}. The validation experiment described in [Leskovec et al., 2007] therefore appears to be limited in its ability to discern a spurious trend from a representative one, as shown in Figure 2.3.

We use the same methodology in combination with other network growth models and graph measures to estimate the impact of various amounts of missing past data (and thus vertex re-discovery). Figure 2.3 shows the output of a single random trial for clarity, but in general, we are interested in the *expected* trend of a measure M in the presence of missing past data, over many random trials. In the next subsection, we describe the stochastic network growth models we will use in our empirical analysis in Section 2.3.3.

2.3.2 Network growth models

A network growth model can be viewed as a probabilistic algorithm that adds nodes and edges to a graph, driven by random noise. A full review of network growth models is beyond the scope of this thesis, but we describe a few key models that will be used here. For a comprehensive survey of network growth models, the reader is referred to [Chakrabarti and Faloutsos, 2006, Chakrabarti et al., 2010].

Definition 2.3.1. *Network growth model.* A network growth model accepts a graph $G = (V, E)$ as input and a vector of parameters Θ, and stochastically produces an output graph $G' = (V', E')$, where $V \subseteq V'$ and $E \subseteq E'$.

Growth models are generally applied recursively, *i.e.,* an initial seed graph is given to the growth model, and the output graph becomes the input graph for the next timestep, for a fixed number of timesteps. Figure 2.4 shows graph layouts of three network growth models for networks of increasing size. The following are three network growth models that we will consider in this chapter:

1. (*Dynamic Random Attachment*) This is possibly the simplest network growth model, originally proposed by Callaway *et al.* [Callaway et al., 2001] to study the properties of randomly growing graphs relative to classical Erdös-Rényi graphs. The growth process is as follows: at every timestep, a new vertex is added to the graph, and with constant probability p, two unconnected vertices are connected with an edge uniformly at random. The observation stream therefore consists of an isolated vertex at each timestep with probability $(1-p)$, or an edge and two or three vertices with probability p. In a variant, we choose an arbitrary pair of vertices for the observation stream instead of an unconnected pair. In the former version, the missing past simulates vertex re-discovery; in the latter, both vertex and edge re-discovery. The first version simulates a citation network, and the latter an interaction network. Finally, since one new vertex is added at each timestep, and the expected number of edges at time t is pt, the expected average degree over time is constant at $2p$.

2. (*Preferential Attachment*) The Barabási-Albert *preferential attachment* (PA) model [Barabási and Albert, 1999, Barabási et al., 2002] was one of the first random graph models intended to generate 'realistic' graphs. It is based on the basic principle that a vertex that has just joined the network will randomly connect to an existing vertex with a probability directly proportional to the degree of the vertex being connected to. While most vertices will end up having a low degree, vertices that initially have a high degree will continue to rapidly increase in degree, as a mathematical embodiment of the 'rich-get-richer' adage. Remarkably, this simple game generates networks with many of the graph-theoretic properties observed in real networks; for example, a skewed degree distribution of vertices, and a small average shortest path length [Bollobás et al., 2001, Newman, 2003, Boccaletti et al., 2006]. Furthermore, the PA model is expected to generate graphs with a slowly growing diameter, *i.e.,* either as $\Theta(\log(V))$ or $O(\log(\log(V)))$ depending on the parameters [Bollobás and Riordan, 2004]. The version of the PA model we consider here was presented as a model of bibliographic co-authorship networks by Barabási *et al.* [Barabási et al., 2002]. Starting with an initial seed graph G_0 and two integer parameters $a > 0$ and $b > 0$, the following describes the simplified PA growth model applied to the graph at each timestep t:

   (a) A new node u is added to the graph and connected to a existing vertices, where the probability of linking to vertex v is defined as

   $$P(u,v) = \frac{N(v)}{\sum_{i \in V} N(i)}$$

   where $N(v)$ is the degree of vertex v.

   (b) b links are created between existing vertices in the graph, where the probability of a link between vertices i and j is defined as

   $$P(i,j) = \frac{N(i)N(j)}{\sum_{s,m \in V, s \neq m} N(s)N(m)}$$

As a variant, in the second step of the algorithm above, if we do not distinguish between already unconnected and connected vertex pairs, the missing past effectively simulates edge as well as vertex re-discovery. Otherwise, only vertex re-discovery is simulated. Note that the version of the algorithm given above would have an expected diameter that grows as $\Theta(\log(V))$ where V is the number of vertices. At each timestep, one new vertex and $(a + b)$ new edges are added to the graph, so the expected average degree over time is constant at $2(a + b)$. If resampling of edges is allowed in the second step, then at most $(a + b)$ new edges are added at each timestep, and the average degree converges from below to $2(a + b)$. Since the average degree converges to a constant, the Densification Power Law is not applicable.

3. (*Forest Fire*) The Forest Fire model was proposed by Leskovec *et al.* [Leskovec et al., 2005] as an alternative to network growth models like Preferential Attachment, which generate graphs with a slowly increasing diameter and constant average degree. Instead, Leskovec *et al.* used the growing network methodology and found that the networks they analyzed showed a rapidly decreasing diameter over time, and superlinearly increasing average degree, which would invalidate Preferential Attachment as a dynamical model for growing networks. They proposed the Forest Fire model to generate graphs with a decreasing diameter and superlinearly increasing average degree.

   Although determining the expected properties of the Forest Fire model appears to be analytically intractable, it is able to generate graphs with the properties mentioned above for specific parameter values described in [Leskovec et al., 2007]. Algorithmically, it can be seen as a form of *copying* model [Kumar et al., 2000], where new nodes pick a set of *ambassadors* and then probabilistically link to their neighborhoods. Given two probability parameters p and r, the graph growth algorithm at each timestep adds a new vertex u as follows [Leskovec et al., 2007]:

   (a) u chooses an *ambassador node* v and links to it.

   (b) Let x and y be two geometrically distributed random numbers with parameters $(1-p)$ and $(1-rp)$. u randomly links to $(x - 1)$ in-link vertices of v and $(y - 1)$ out-link vertices of v, excluding any that have already been visited in the current iteration.[4]

   (c) The second step repeats at all nodes that have just been linked to, until the process dies out.

Consider the output of iterating any of the network growth models above. Recall that we start with an initial seed graph, possibly empty, which is designated $G_0 = (V_0, E_0)$. At each timestep, the network growth model generates an *observation graph* $G' = (V', E')$, which is added to the input graph to produce the output. Note that V' and E' can have elements in common with the input graph, but are not necessarily strict supersets or subsets of it.

Let Grow be such a network growth model:

$$\textsc{Grow} : G \times \Theta \rightarrow G'$$

where G is the input graph, Θ is set of model parameters, and G' is the output observation graph describing changes made to the input graph. Let C be the number of initial timesteps to censor to simulate the effect

[4]The -1 constants on x and y are not mentioned in the description of the algorithm, but are critical for reproducing the results in [Leskovec et al., 2005, Leskovec et al., 2007].

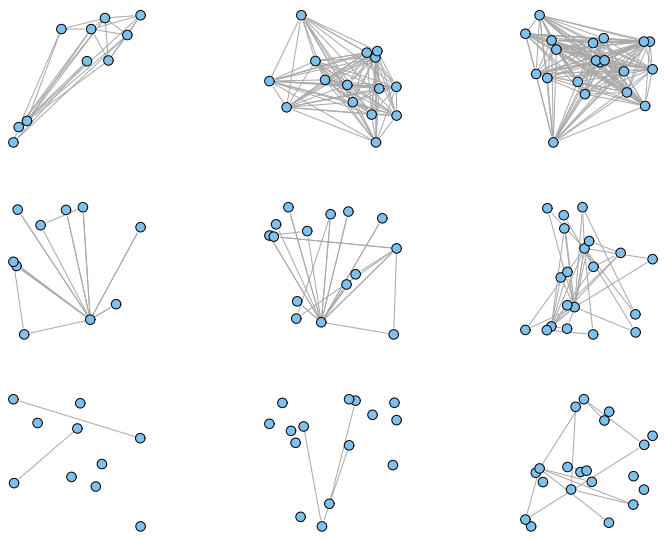


Figure 2.4: Examples of the output of network growth models for graphs of 10, 15, and 20 nodes: forest-fire (top row), preferential attachment (middle), random attachment (bottom).

of re-discovery. Starting with timestep 0 and a seed graph G, we iterate the network growth model in the following manner to generate the ground truth network:

$$O_1 = \textsc{Grow}(G, \Theta) \qquad G_1 = G \cup O_1$$
$$O_2 = \textsc{Grow}(G_1, \Theta) \qquad G_2 = G_1 \cup O_2$$
$$...$$

This yields the underlying (ground-truth) network:

$$\langle \mathbf{G} \rangle = \langle G, G_1, ... \rangle$$

We then censor the first C timesteps to generate the observed, aggregated network $\langle G^+ \rangle$ (see Definition 2.1.2).

$$\langle G^+ \rangle = \langle O_C, O_C \cup O_{C+1}, O_C \cup O_{C+1} \cup O_{C+2}, ... \rangle$$

Depending on the model, the occurrence of some vertices in $O_t, t \geq C$ will be caused by re-discovering a pre-existing vertex. This could happen, for example, when a new edge connects to a vertex that exists in the censored part of the network. In some network models, existing edges can be re-activated, so there is also the possibility than an edge in some O_t for $t \geq C$ is a re-activation of an edge in the censored part of the network. This describes a very common 'missing-past' scenario in network data collection, and it is important to keep in mind that the properties of interest are those of the underlying network $\langle \mathbf{G} \rangle$.

2.3.3 Empirical results

The experimental methodology for simulating the missing past is simple: iterate a network growth model for C timesteps to generate a ground truth network, and then 'fork' its evolution into a secondary, "missing past" network that only receives updates to the ground truth network from timestep C onwards.[5] For example, if the ground truth network receives edge (u, v) at timestep $C + 1$, the missing past network at the same timestep contains *only* the edge (u, v). The larger the value of C, the larger the missing past network, and depending on the graph model, the higher the prevalence of the re-discovery process in uncovering censored vertices. We might intuitively expect that for small values of C, any difference between the properties of the ground truth and missing past network would converge to the same value very quickly. By varying C, we can test this assumption with any network growth model.

For each of the three network growth models described earlier, we started with the simplest seed graph permissible by the model, generally a single isolated vertex. The growth model was iterated for a total of 1,000 timesteps, with the missing past size ranging from 50 to 250 timesteps in increments of 50 timesteps. Note that this represents a very small amount of missing past data. In all three models, this equates to a missing past graph of just 50 vertices. Various graph theoretic properties were measured on both the missing past and ground truth networks every 10 timesteps. All properties were averaged over 500 random trials. Shortest paths were computed exhaustively between all reachable vertex pairs using the `igraph` network library. We present results in the following subsections for the following model configurations:

1. Random attachment with edge creation probability $p = 0.8$ (sparse network[6]).
2. Preferential attachment with $a = 2$ and $b = 2$ (similar to [Barabási et al., 2002]).
3. Forest fire with $p = 0.35$ and $r = 0.57$ ('sparse graph' instance in [Leskovec et al., 2007]).

We report trends in the average shortest path length, effective diameter, clustering coefficient, largest eigenvalue of the adjacency matrix, and the densification exponent of the DPL.

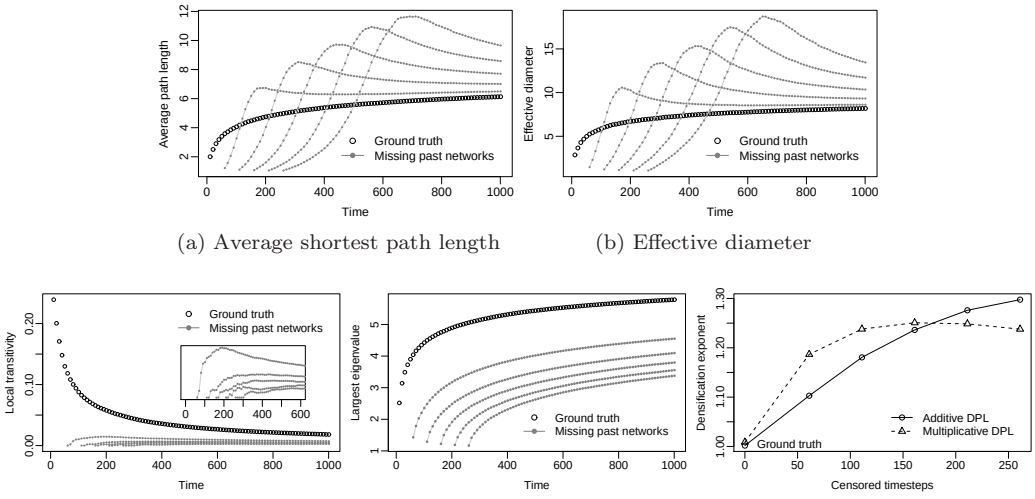
Random Attachment

Figure 2.5 shows various properties of the ground truth network relative to 5 missing past networks. We focus first on the average shortest path length and the effective diameter, which are both statistics of the pairwise shortest path length distribution. Even with a relatively small amount of missing data of 50 timesteps, the trend in the observed dataset is always the opposite of the trend in the underlying network. This is not only a qualitative difference (increasing vs. decreasing), but also numerically quite a large difference at short times.

Predictably, as the amount of missing data increases, the observed trend takes longer to converge to the true trend; with 100 censored timesteps (the second gray trend from the left), the ground truth and missing past trends appear to be converging in about 1,000 timesteps – which represents more than 10 times the number of vertices in the missing past. This convergence takes considerably longer with more missing

[5] In Unix-like operating systems, the $fork()$ system call creates a second concurrent copy of a process. The difference is that unlike a Unix process, our 'missing past' network does not inherit any data from the original ground truth network.

[6] Note that the edge creation probability is distinct from the edge probability in classical Erdös-Rényi graphs. In the random attachment growth model, $\lim_{p \to 1} E(t) = V(t)$, which is the classical definition of a sparse graph.



(a) Average shortest path length

(b) Effective diameter

(c) Clustering coefficient (inset shows early stages of missing past network)

(d) Principal eigenvalue

(e) Densification exponent

Figure 2.5: Random attachment network model with $p = 0.8$: the effect of 5 different amounts of missing data on the ground truth (empty circles) and missing past (filled circles) networks.

data. For example, with 250 missing past timesteps (and thus, 250 missing past vertices and $250 * 0.8 = 200$ edges in expectation), the missing past trend in the average shortest path length is a little less than double its ground truth value at the end of 1000 timesteps. The effective diameter is almost three times its true value. Perhaps what is of most concern is that qualitatively, the missing past trends appear to be decreasing, whereas the ground truth trend is increasing in both cases. The clustering coefficient and principal eigenvalue appear to be good qualitative approximations even with missing past data. However, the former is a biased measure that is intrinsically correlated with the average degree, and the latter is known to correlate with the maximum degree, decreasing their value as dynamic graph characterizations.

Finally, the most surprising finding here is that of the DPL densification exponent α (see Definition 2.1.6). Recall that α lies strictly between 1 and 2, and that a value greater than 1 indicates that the number of edges is growing superlinearly relative to the number of nodes. In the random attachment model, the number of edges is a constant function of the number of nodes, so the DPL is known not to hold in this case. However, the re-discovery process has the following effect in the early stages of observed network growth: when a new edge is created between two vertices in the censored graph, the observations show the appearance of two new vertices and one new edge, instead of one new edge and no new vertices. The number of edges therefore appear to grow slowly initially, as censored vertices are re-discovered. This manifests in an apparently super-linear growth of the number of edges compared to the number of nodes over the later portion of the observation period, whereas the bias is really towards *sub*-linear growth in the early stages. Unfortunately, both phenomena can manifest as super-linear fits with regression.

For each missing past network, we used non-linear least-squares regression to fit the Additive DPL equation and linear least-squares on log-transformed data to fit the Multiplicative DPL equation. Figure 2.5e shows the densification exponent α for each missing past dataset, with $C = 0$ corresponding to the ground

truth dataset. In all cases, α is greater than 1, apparently suggesting that densification is taking place in the underlying system when we know that this is not the case. Larger amounts of missing past lead to higher densification exponents using the Additive DPL equation; the Multiplicative DPL appears to level off and then decrease. This can be explained by the re-discovery process being more likely to uncover nodes that already existed in the censored portion of the network due to its increased size.

Preferential Attachment

We now consider a network that is growing according to the Preferential Attachment model, which is a more realistic than the random attachment model analyzed in the previous section. We conduct a similar set of experiments as with the random attachment model by censoring a portion of the initial output of the network growth model.

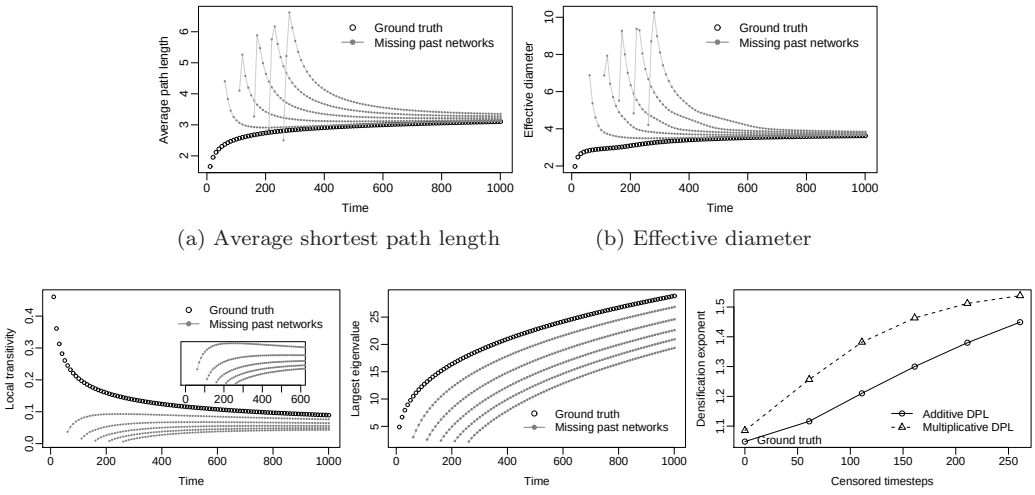
The results are quite similar to those of the random attachment model. Figure 2.6 shows measurements for the ground truth and missing past networks of various sizes. As with the random attachment model, the missing past networks exhibit false trends in both the average shortest path length and effective diameter over time. However, they are significantly more pronounced than in the random attachment model, both qualitatively as well as quantitatively. Although the trend in the underlying networks is growing slowly, as predicted by theoretical results [Bollobás and Riordan, 2004], the observed network exhibits an initially sharply decreasing effective diameter, followed by slow convergence to the true trend. Thus, we can conclude that the spurious trends in the random attachment model were not solely the result of the random network structure. The clustering coefficient measure also exhibits a false trend at short times, but converges to the true trend quickly.

The DPL densification exponents α also show strong variability with the amount of missing past data. Particularly, the Additive DPL equation appears to exhibit a linear dependence on the size of the missing past network. However, it is unlikely that this property, even if it holds generally, can be used to determine the size of the missing past from the densification exponent; that would only be applicable if the underlying network was truly growing according to preferential attachment. Once again, two sources of concern are the strong dependence of α on the missing past size, and the very fact that α indicates densification with a very small amount of missing past when the ground truth network does not possess that property.

Forest fire

Finally, we analyze a configuration suggested by the authors of the Forest Fire model. Since the model appears to be analytically intractable [Leskovec et al., 2007], we use a parameter set that has been empirically shown to generate a growing sparse network with a slowly increasing effective diameter. We also adjusted the timescale used for the experiments to match that in the original paper to allow for an easy comparison of behavior; it is about 10 times longer than the timescale used for the other experiments. Since the Forest Fire model is a type of 'copying' model [Kumar et al., 2000], we might expect different results than the random and preferential attachment models. Figure 2.7 shows results for the 'sparse' configuration of the Forest Fire model.

The trends in path length statistics are somewhat unusual: the ground truth network has a slowly increasing diameter, much like the preferential and random attachment models, the effective diameter in the missing past networks in Figure 2.7b are essentially constant after a brief initial period. This can be



(a) Average shortest path length

(b) Effective diameter

(c) Clustering coefficient (inset shows early stages of missing past network)

(d) Principal eigenvalue

(e) Densification exponent

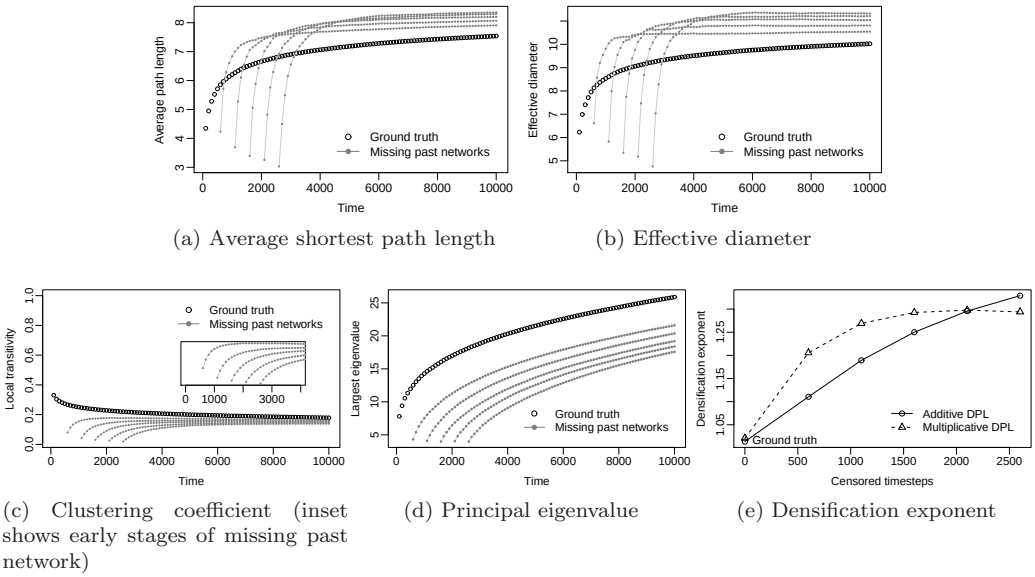
Figure 2.6: Preferential attachment network model: the effect of 5 different amounts of missing data on the ground truth (empty circles) and missing past (filled circles) networks.

explained as a facet of the forest fire graph generation process. When a vertex is revealed in the missing past network, it burns edges and thus re-discovers old vertices in the ground truth network. In practice, this 'forest fire' process (described in Section 2.3.2) burns out very quickly, and thus the re-discovered vertices are short distances away from the new vertex, keeping the effective diameter low. When the network is large enough, this effect becomes insignificant, and the trend manifests as an almost constant effective diameter. This is therefore a case when the data suggests a steady state where there is none.

Other properties not based on shortest paths display errors or consistency similar to the random and preferential attachment models: the clustering coefficient at short times shows a trend contrary to the ground truth, and the principal eigenvalue tracks its ground truth value qualitatively, if not quantitatively. The densification exponent, however, shows the same dependence upon the amount of missing past as the other models, ranging between approximately 1 and 1.3 for a ground truth value near 1. We note that the dependence of α resembles the trend in the random attachment model (Figure 2.5), suggesting that the dependence of the bias in certain cases might be related to the relative sizes of the missing past and observed networks. Whether this is a universal feature, and if the size of the missing past network might be computable from it, is a question for future research.

2.4 Sensitivity of measured trends in interaction networks

In the previous section, we asked how sensitive certain measures are on growing *citation* networks, given the presence of various amounts of missing past data. In this section, we ask if the same trends in the measures of growing *interaction* networks are meaningful. Recall that in an interaction network there is an underlying dynamic process that causes entities to interact with each other over time, and that we discover



(a) Average shortest path length

(b) Effective diameter

(c) Clustering coefficient (inset shows early stages of missing past network)

(d) Principal eigenvalue

(e) Densification exponent

Figure 2.7: Forest fire model with $p = 0.35$ and $r = 0.57$: the effect of 5 different amounts of missing data on the ground truth (empty circles) and missing past (filled circles) networks.

the structure of the network only through observing these interactions. This is an appropriate model for many communications and information networks, such as phone call, e-mail, physical proximity, and instant message logs. For example, in the case of e-mail networks, if a user never sends a single e-mail, their existence will generally not be noted on transmission logs, and conversely a user's first*observed* e-mail transmission would suggest true network growth to the observer, regardless of whether the observation contains the first instance of that edge.

One of the assumptions in measuring the change in a growing interaction network is, naturally, that the underlying network is truly changing with respect to some measure M. If this were not the case, then any trend in M over time would simply be a reflection of the convergence of a sampling process to the limiting value of M. It would be a reflection of the sampling process, a product of interaction network dynamics and missing past data. In fact, at least one study uses the method described here to study the convergence of various network measures to limiting values [Latapy and Magnien, 2008]. In this section, we expand upon an elegant conceptual framework proposed by [Pedarsani et al., 2008] called *edge sampling* to ask if a given dynamic network dataset is growing at all. In other words, is there a plausible dynamic sampling process operating on an *unchanging* network that can statistically explain a trend in measure M when M is truly constant?

Definition 2.4.1. *(Edge Sampling)* The basic assumptions of edge sampling are as follows:

1. There is a fixed or slowly changing (relative to the size of observations) underlying graph.
2. Edges in this underlying graph randomly 'activate' according to a sampling distribution at each timestep, and are thus discovered by the observer.

Since we are assuming a steady state with respect to measure M, under a consistent sampling process

and enough data, we can treat the final aggregate static network obtained from a sequence of observations as a good approximation of the underlying system. We then specify an *edge sampling* model to randomly activate edges from the aggregate graph at each timestep of observations, simulating interactions (and thus observations) along an unchanging graph. By starting with a real dataset and specifying an edge sampling model, we can perform many Monte Carlo edge sampling simulations and determine the expected trend over time in any graph theoretic property, as well as its sampling distribution, as a form of statistical permutation test [Good and Wang, 2005]. Although extremely computer intensive when operating on large networks, permutation tests have a long history and are extremely powerful analytical tools.

In the next subsection, we describe the basic edge sampling model proposed in [Pedarsani et al., 2008]. In Section 2.4.2, we describe four additional edge sampling models to satisfy the second part of Definition 2.4.1. Since permutation methods are very computationally intensive, we report results on a number of smaller datasets:

1. *Enron (internal).* All e-mail traffic between `@enron.com` e-mail addresses, quantized by month.
2. *IMDB Photos.* Celebrities photographer together, quantized by month.
3. *DBLP-FOCS.* A small sample of the DBLP digital bibliographic database, focusing on co-authorship in the Foundations of Computer Science conference. Note that the earlier years have significant missing data issues. The quantization timestep is one year.
4. *HEP-Th.* Citations in high-energy physics publications, quantized by month.

2.4.1 Uniform Edge Sampling

The basic edge sampling model assumes an underlying graph $G = (V, E)$, from which edges are 'activated' at each timestep uniformly and independently at random with a fixed probability p_e [Pedarsani et al., 2008]. Vertices are only discovered as a consequence of an edge activating; thus, isolated vertices of degree 0 are never discovered. We present a minor modification of the edge sampling model to a temporal setting here.

Let $v(t)$ and $e(t)$ be the number of nodes and edges that have been discovered from the underlying graph at time t, with $v(0) = 0$ and $e(0) = 0$. At each timestep $t \geq 1$, each edge in the underlying graph activates with fixed probability p_e independent of all other edges, and is therefore discovered if it has not previously activated. The number of activations of an edge over t timesteps is therefore binomially distributed as $Bin(t, p_e)$, and the probability that the edge does not activate at all after t timesteps (and thus remains undiscovered) is q_e:

$$q_e = (1 - p_e)^t$$

At time t, the expected number of discovered edges $e(t)$ is the mean of a second binomial distribution with probability $(1 - q_e)$ for E trials.

$$\begin{aligned} \mathbf{E}[e(t)] &= E \cdot (1 - q_e) \\ &= (1 - (1 - p_e)^t) \\ Var[e(t)] &= E \cdot (1 - (1 - p_e)^t)(1 - p_e^t) \end{aligned}$$

A node is discovered if one of its adjacent edges fires, and remains undiscovered if all of its edges do not fire. The probability of a node of degree d remaining undiscovered at each timestep is therefore q_e^d. After t timesteps, this probability is $(q_e^d)^t$. Conditioning on node degree, where $f_D(d)$ is the proportion of nodes in the graph with degree d, the probability of a node not being discovered after t timesteps is given by:

$$q_v(t) = \sum_{d=1}^{V-1} q_e^{d \cdot t} \cdot f_D(d)$$

Thus, the expected number of nodes $v(t)$ at time t is:

$$\begin{aligned} \mathbf{E}[v(t)] &= V \cdot (1 - q_v(t)) \\ &= V \cdot (1 - \sum_{d=1}^{V-1} q_e^{d \cdot t} \cdot f_D(d)) \end{aligned}$$

Given the size (number of vertices and edges) and empirical degree distribution of an 'underlying' network, and a fixed sampling probability p_e, the edge sampling model lets us analytically determine the number of vertices and edges we can expect to see at each timestep in a dynamic discovery process. This, in turn, allows us to theoretically determine if we can expect to see edge densification for a given value of the sampling parameter p_e. It was shown that this model leads to densification in graphs with a power-law-like distribution [Pedarsani et al., 2008]. Analytically determining more complex properties, such as the average shortest path length, quickly becomes difficult[7], but can be investigated using Monte Carlo analysis.

2.4.2 Other edge sampling models

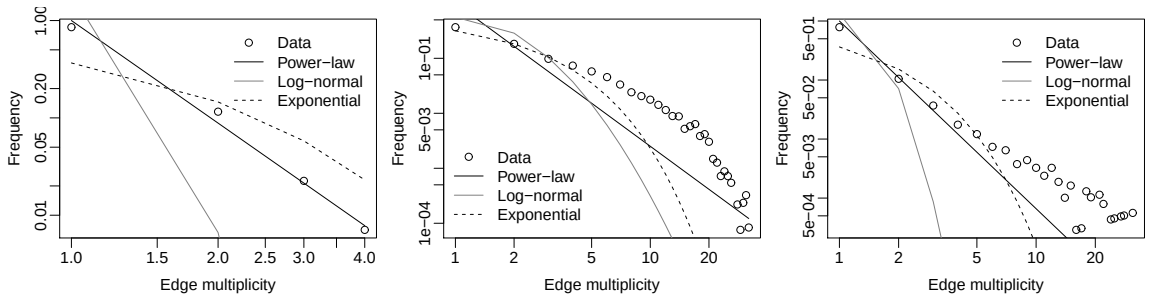
We describe four additional edge sampling models to be used for hypothesis testing. Each model attempts to preserve some statistic of the observed dataset.

Size-preserving edge sampling

The *size-preserving* (SP) edge sampling model samples $|E_t|$ edges from the static graph at each timestep, independently and uniformly at random without replacement, where $|E_t|$ is the number of edges actually observed in the dataset at time t. This model effectively stipulates that there is an unknown process governing the number of edges observed at each timestep (an observation or sampling process), but that within each timestep, edges are activated uniformly at random from the underlying network (the process governing interactions in the underlying system). Note that edges are chosen without replacement for generating the observation for a particular timestep, but with replcement across timesteps.

Figure 2.8 shows the distribution of edge multiplicity values (*i.e.,* , the distribution of the number of times each edge was observed in the dataset) for three datasets that exhibit edge multiplicity. A notable feature of all three distributions is their heavy skew: most edges are observed only once within the observation period, but a significant number are observed multiple times. In the same figure, we also show fitted distributions

[7]As an example, the proof techniques used to analytically determine the diameter of a scale-free graph are far from trivial [Bollobás and Riordan, 2004].



(a) DBLP-FOCS (b) Enron internal (c) IMDB Photos

Figure 2.8: Distribution of edge multiplicity (also known as *support*) in various datasets, along with several common fitted distributions, shown on a doubly logarithmic scale. Although the data is heavily skewed, none of the distributions appear to be a good fit.

to the data: a power-law fit using the methods described by [Clauset et al., 2009], and exponential and log-normal distributions fitted using maximum likelihood estimators.

In the SP edge sampling model, each edge has the same probability of being chosen at a particular timestep, although with different probabilities across timesteps, so the distribution of edge multiplicities is Poisson-binomial [Wang, 1993], which is a general case of the binomial distribution. If E is the number of edges in the aggregate static network, and E_t the number of edges observed in the dataset at time t, then the probability of picking an edge at any timestep is E_t/E. Since this is generally small relative to the number of timesteps, the Poisson-binomial distribution is well-approximated by a Poisson distribution [Chen and Liu, 1997], which can appear as a skewed distribution qualitatively similar to the empirical distributions in Figure 2.8. Although a Poisson distribution does not yield a good fit to the data in Figure 2.8, it can serve as a useful first approximation.

Degree-preferential edge sampling

Similar to the preferential attachment growth model [Newman, 2001b], we propose the *degree-preferential* (DP) discovery model as a refinement of the SP edge sampling model. In the DP model, E_t edges are still picked at each timestep, but not uniformly at random. Instead, the probability of an edge being sampled is directly proportional to the product of the degrees of its vertices, *i.e.,* edges between nodes prolific in connections are more likely to be re-activated. Specifically, the probability of sampling an edge $e = (u, v)$ at any timestep is given by:

$$p(e_{uv}) = \frac{d(u) \cdot d(v)}{\sum_{i \neq j} d(i) \cdot d(j)} \tag{2.3}$$

where $d(i)$ is the degree of vertex i. Algorithmically, sampling E_t edges without replacement from a graph according to Equation 2.3 can be carried out efficiently using the weighted random reservoir sampling (WRS) algorithm described in [Efraimidis and Spirakis, 2006].

Intuitively, the DP model still stipulates some unknown process dictating the number of edges observed at each timestep, but a slightly different underlying interacting process that governs edge activations. In the DP model, being connected to a high degree vertex increases a vertex's chance of generating an activation at

any given timestep. Furthermore, interactions between two connected high-degree vertices are more likely than between a pair of low-degree vertices.

Rate-preserving edge sampling

The *rate-preserving* (RP) edge sampling model samples each edge independently at each timestep with probability equal to its estimated probability from data, *i.e.,* the edge multiplicity divided by the total number of timesteps. This is perhaps the most realistic of the edge sampling models presented here. Although it makes a strong independence assumption, it reproduces a number of statistics of the original dataset, such as the distribution of edge multiplicities. For an edge with multiplicity c in the dataset, its multiplicity under the RP model is a binomially distributed random variable with success probability c/t.

Size and Count-preserving edge sampling

Similar to the size-preserving sampling model, the *size and count-preserving* (SCP) sampling model preserves not just the number of edges observed at each timestep in the original dataset, but also the total number of activations of each edge, *i.e.,* edge multiplicity. Unlike the RP model, however, the multiplicity of each edge is a constant c instead of a random variable binomially distributed with mean c.

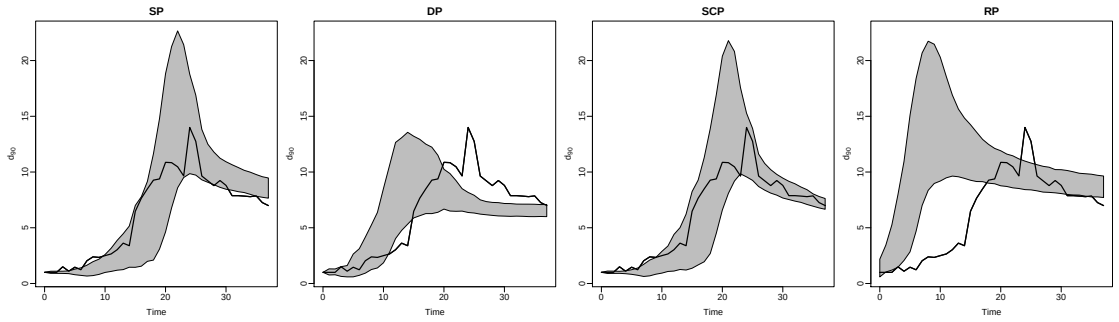
Preserving both the number of edges at each timestep as well as the total count of each edge is a non-trivial problem. We use a common Markov Chain Monte Carlo (MCMC) approach to perform this randomization, which has been previously used to randomize single graphs while preserving properties like the average path length [Hanhijärvi et al., 2009], and originally developed to generate random bipartite perfect matchings [Broder, 1986].

The approach involves representing each possible satisfying dynamic network as a node in a Markov chain with equiprobable transitions to all nodes that can be reached by performing a single *swap* operation. In our context, this operation swaps two distinct edges in different timesteps such that the total count of edges in each timestep remains the same before and after the swap. By iterating this Markov chain, we eventually reach the limiting uniform distribution and have successfully randomized the network subject to the size- and count-preserving constraint.

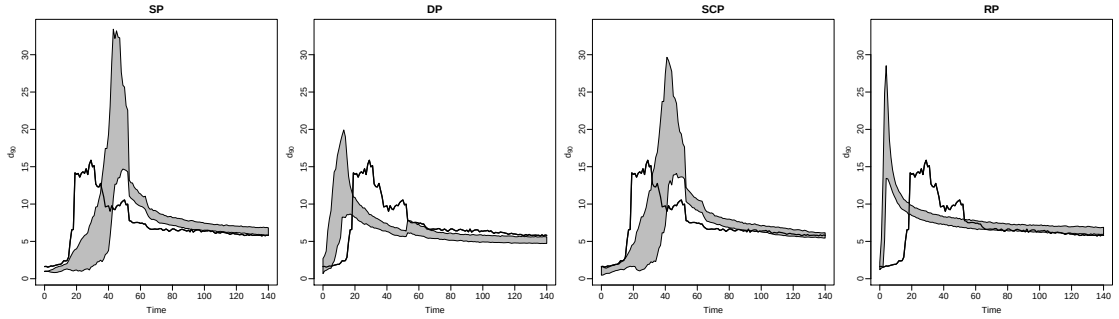
In practice, the Markov chain is iterated until it mixes before being used, but it is difficult in general to tell when the limiting distribution has been reached; in [Hanhijärvi et al., 2009], a constant number of iterations was used. We run the chain forward for a number of iterations equal to the total number of edge occurrences before using the randomized network, and then use each randomized network as the starting point for subsequent randomizations.

2.4.3 Empirical results

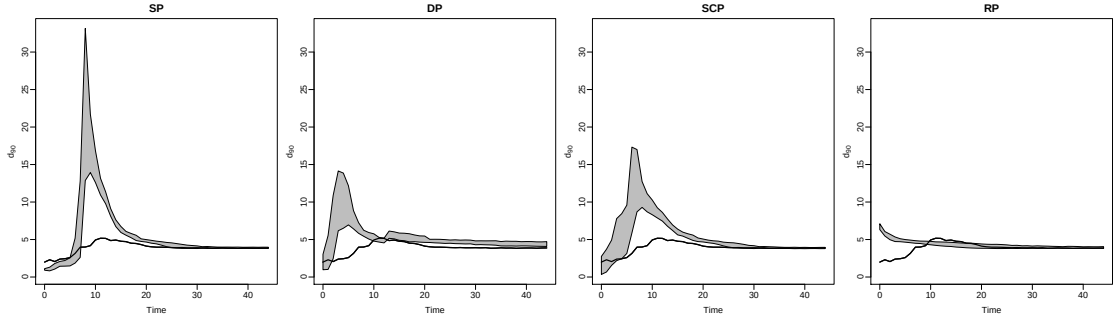
Recall the that the four discovery models we proposed were: rate-preserving (RP), size-preserving (SP), degree-preferential (DP), and size- and count-preserving (SCP). Given the original dataset, none of these edge sampling models require any parameters. We ran each model between hundreds and thousands of times (depending on the size of the dataset) on each of the five datasets described earlier. For each run, we sampled edges from the final aggregate static network for the same number of timesteps as the original dataset, and recorded the values of various graph properties at each timestep. From these values, we can estimate the



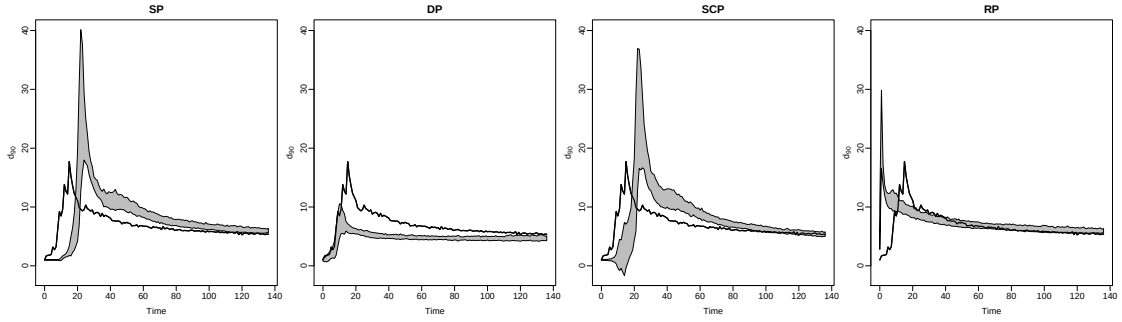
(a) DBLP-FOCS



(b) IMDB Photos



(c) Enron



(d) HEP-Th

Figure 2.9: The effective diameter over time of various datasets (solid line), and a band of two standard deviations around the expected trend under each edge sampling model.

mean property value at each timestep under a given edge sampling model, and thus the mean trend, as well as the variance and other statistics. Intuitively, if the edge sampling models are bad fits, the mean trend will be far from the actual observed trend.

Figure 2.9 shows the mean effective diameter d_{90} of each edge sampling model, along with a band showing two standard deviations. Closed circles represent the original value of the effective diameter measured in the dataset at each timestep. The DBLP-FOCS dataset, shown in Figure 2.9a, is a good example of a reasonable fit to an edge sampling model. The SP and SCP appear to be very good fits to the observed data. In both cases, the qualitative trend from the edge sampling model follows the trend observed in the data quite closely. Furthermore, points in original dataset are quite often within two standard deviations of the edge sampling trend. The DP model yields a slightly worse fit, with the d_{90} value ending higher than the observed data. Fits for the other datasets are not as close as the DBLP-FOCS dataset, but all show the same qualitative trend of a decreasing effective diameter.

Recall that one interpretation of the edge sampling model is that the underlying network is not changing at all; edges are merely being activated at each timestep and discovered. This is exactly the algorithm used for the sampling process – edges are sampled from a fixed, unchanging, underlying network. In spite of this, the trend suggested by the measurements is that the effective diameter in the underlying network is decreasing. The spurious trends are not insignificant either, and all follow the same decreasing trend, which might suggest a universal phenomenon when it is demonstrably an artifact.

2.5 Summary and suggestions

In this chapter, we described two common methods for measuring the properties of dynamic networks over time, and summarized the trends in common properties reported on multiple datasets. We also classified dynamic network data into two broad classes: *citation* networks and *interaction* networks, depending on whether the change in the structure of the underlying process is either directly observed or measured through a proxy of interactions occurring on the network, with possibly independent dynamics. Most importantly, we systematically analyzed whether trends in common measures can be attributed to true change in the underlying system, or explained as either missing past artifacts (citation networks) or artifacts introduced by the dynamics of interactions (interaction networks).

There is an inherent bias in the growing network methodology for citation network measurement at short times, particularly in networks with a missing past. This bias can be understood as being caused by two sampling processes at play in the collection of network data: *re-discovery* of pre-existing vertices, and true *growth* in the underlying network. Prior studies assumed that all observations could be attributed to growth of the underlying network. In most cases, however, the sampling process is doubly stochastic as explained, with both discovery and growth playing a role. Very small amounts of missing past data can manifest as significantly erroneous trends in observations. Using synthetic data generated by well-known random graph models, we were able to show that: *densification (super-linear growth of edges relative to nodes) can manifest in networks that are not densifying, effective diameter can manifest as sharply decreasing in networks with a slowly increasing effective diameter, and these biases cannot be detected by withholding a portion of the observations.*

Similarly, trends in common measures in a number of real interaction network datasets can be explained

using simple random models of interactions occurring along an unseen graph structure that we want to measure. The change in the graph is not directly observable, but we assume that the observation of an interaction indicates the existence of an edge between the adjacent vertices. New interactions occurring along freshly created paths eventually reveal those paths to the observer, but conversely, there is no way to tell if a newly discovered path is actually new. Thus, we hypothesize that there might be little or no growth in the underlying network, and the observations are purely a product of discovering pre-existing vertices and edges. Using randomization tests, the expected trend under this hypothesis matched real data qualitatively, and in many cases, quantitively as well, within acceptable statistical bounds. Using the growing network method on an interaction network therefore reveals a trend that can also be interpreted as the convergence of measure M to its limiting value under a steady state hypothesis for M.

In the rest of this section, we focus on suggestions for future research.

2.5.1 Choosing an appropriate network representation

Recall that we had previously been able to classify network datasets into two categories: *interaction* networks, where the observations at any timestep represent instantaneous associations between vertices that can re-occur in the future, and *citation* networks, where observations represent the one-time formation of permanent links. We deal with each separately:

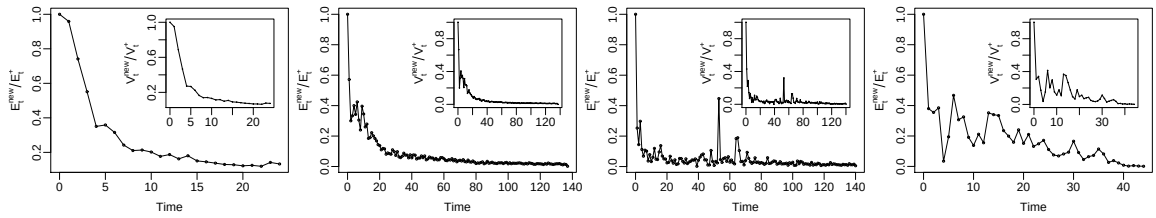
- *(Interaction networks)* A fully dynamic or dynamic interaction network with some form of smoothing or vertex and edge decay might be the most appropriate representation (see Section 2.1). If vertex or edge removal is a characteristic of the underlying system, then one of the following two methods could be used to reduce noise in the time series of observations:

  1. *(Weighting)* Use a decay function to weight edges, and then either use weighted graph measures [Newman, 2001a, Barrat et al., 2004, Newman, 2004, Zhang and Horvath, 2005, Barthelemy et al., 2005], or remove vertices and edges that have fallen below a threshold weight at any time[8] [Kossinets and Watts, 2006, De Choudhury et al., 2010].

  2. *(Smoothing)* An alternative to weighting graphs is to expand the timescale to consist of larger timesteps in order to reduce noise. For example, if the observations of a network are at the resolution of a day, one could consider a dynamic network that aggregates a month of observations into a single timestep. Furthermore, one could consider a sliding window over this observation stream instead of a shifting-window quantization. A number of studies have looked at the effect of different timescales on network properties [Delvenne et al., 2010, De Choudhury et al., 2010, Eagle and Pentland, 2006], and there have been at least two algorithms proposed to find meaningful timescales for a time series of graphs [Sun et al., 2007, Sulo et al., 2010].

- *(Citation networks)* Although growing networks can be an appropriate representation for citation networks, the missing past issue introduces a vertex discovery process. In this case, we can only suggest that the network is allowed to stabilize to the point where new observations comprise a smaller portion of the aggregated network. This eliminates the noisy initial portion of the network when the

[8] A recent arXiv working paper suggests that this method might be too noisy in general [Thomas and Blitzstein, 2011].



(a) Patent citations (b) HEP-Th co-authorship (c) IMDB Photos (d) Enron internal

Figure 2.10: The number of previously unobserved edges and vertices (insets) at each timestep relative to the size of the aggregated network at that timestep for four real datasets.

discovery process is most likely to have a significant impact. Note that taking such a precaution is a heuristic; it does not solve the missing past problem.

Figure 2.10 shows the proportion of new edges and vertices introduced at each timestep as a function of the aggregate growing graph size at that timestep. As expected, new vertices and edges comprise a smaller portion of the aggregate graph as time progresses, so a simple heuristic might be to wait until new observations comprise no more than a small portion of the aggregate graph before considering measurements. This is similar in principle to the initial 'burn-in' period of many iterative statistical algorithms, where the output of an algorithm is discarded until the underlying model has reached stationarity [Hastie et al., 2001, p.280]. Setting this threshold arbitrarily at 5% would entail dropping approximately the first 20 measured data points of the patent citations dataset [Leskovec et al., 2007], the first 40 timesteps of the HEP-Th co-authorship dataset [Leskovec et al., 2007], and the first 20 timesteps of the IMDB photos dataset [Lahiri and Berger-Wolf, 2008].

An entirely different solution would be to abandon the graph-theoretic representation of networks for one that is perhaps less sensitive to noise and missing past effects. As we mentioned earlier, spectral density plots have been proposed as a concise and effective way to look into the structure of networks [Banerjee and Jost, 2009], but doing so over time poses both a mathematical and visualization challenge. Alternatively, latent space models embed the nodes of a social network into a vector space, using edges to define some form of distance function [Hoff et al., 2002]. These models have been extended to dynamic networks as well [Sarkar and Moore, 2005], but represent a fundamentally different direction than classical graph theory.

2.5.2 Equilibrium assumption

Recall that a fundamental assumption in static network analysis is that the physical system is in equilibrium with respect to graph theoretic properties, so taking a large enough sample of the network yields representative approximations to the properties of the underlying system. On the other hand, one of the tenets of growing network analysis is that the properties of the underlying system are changing over time, as evidenced by the trends in graph theoretic properties over time.

As a telling example of this difference, consider two lines of research that use the same methodology, report essentially the same observations, but start with different assumptions and come to very different conclusions. Latapy and Magnien [Latapy and Magnien, 2006] ask how large a network sample must grow until the properties of the aggregated network reach a steady state. Their methodology samples some

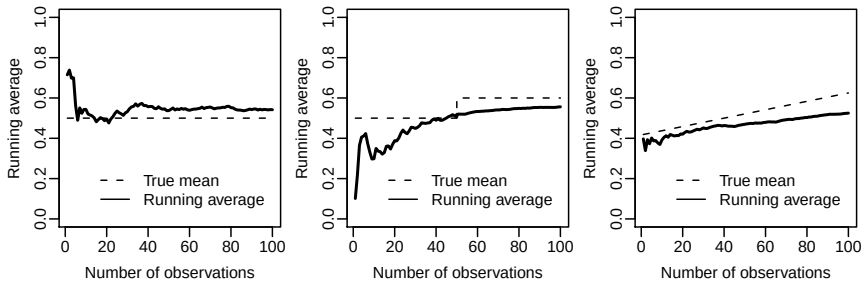


Figure 2.11: An illustration of the running average of a growing pool of random variates drawn from various uniform random distributions, as an analogy to the difficulty of detecting nonstationarity with the growing network methodology at short times. The task is akin to inferring the behavior, not the value, of the dotted line (ground truth) from a single growing pool of observations (solid line).

networks over time, only because this is a natural order for those datasets, and shows graph-theoretic measurements that eventually converge to a fixed value. For measurements that converge, the authors conclude that the sampling process was effective and the steady-state values of the underlying network have been discovered. On the other hand, classic studies of growing networks use the same methodology to infer how the properties of the underlying network are *changing* over time [Barabási et al., 2002, Leskovec et al., 2005]. In both cases, we see qualitatively identical plots of the same property (*e.g.,* average shortest path length over time in an aggregated network), but diametrically opposite inferences about the underlying network.

So are our observations successively converging to the underlying network, which is in a steady state, or are each of our observations representative enough to allow us to infer that the underlying network is changing? There is unlikely to be a general answer to this question, since one can imagine different situations in which each is plausible. The methods we have presented earlier in this chapter allow us to explicitly construct some of these situations. Thus, the equilibrium assumption, or lack thereof, needs to be justified. Consider the numerical analogy to growing network analysis shown in Figure 2.11. In each figure, a process is emitting uniformly distributed random numbers, either with a constant expectation or a trend as shown by the dotted line. The solid line shows the running average of the observations, similar to how growing networks accumulate observations into a single graph. The solid line at each time point represents the best estimate for the expectation of the process. In an alternative interpretation of the solid line, the expectation of the underlying process is either roughly constant (leftmost figure), or slowly growing (middle and right figures). However, the ambiguity of the setup and the actually measured solid line makes it difficult to infer the behavior of the dotted line. Neither is correct in all cases, so the observed measurements themselves cannot be used as a justification for assuming equilibrium or non-equilibrium at short times.

2.5.3 Dynamic measures

Finally, we note that interaction networks appear to be more common than citation networks. Since interaction dynamics can play a big part in our view of the underlying physical system, it might be advantageous to develop measures that directly extract information about the dynamics of the process, rather than measuring properties of a static structure at fixed time intervals. This is the motivation for the techniques developed

in the remaining chapters of this thesis.

Bibliography

[Acar et al., 2009] Acar, E., Dunlavy, D., and Kolda, T. (2009). Link prediction on evolving data using matrix and tensor factorizations. In *IEEE Intl. Conf. on Data Mining Wkshps.*, pages 262–269. IEEE.

[Adar and Adamic, 2005] Adar, E. and Adamic, L. (2005). Tracking information epidemics in blogspace. In *Proc. of the 2005 IEEE/WIC/ACM Intl. Conf. on Web Intelligence*, pages 207–214. IEEE Computer Society.

[Ahn et al., 2007] Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., and Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. In *Proc. of the 16th Intl. Conf. on World Wide Web*, pages 835–844, New York, NY, USA. ACM.

[Akoglu and Faloutsos, 2009] Akoglu, L. and Faloutsos, C. (2009). RTG: a recursive realistic graph generator using random typing. *Machine Learning and Knowledge Discovery in Databases*, pages 13–28.

[Akoglu et al., 2008] Akoglu, L., McGlohon, M., and Faloutsos, C. (2008). RTM: Laws and a recursive generator for weighted time-evolving graphs. In *Proc. of the 8th IEEE Intl. Conf. on Data Mining*, pages 701–706. IEEE.

[Albert and Barabási, 2002] Albert, R. and Barabási, A. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47–97.

[Albert et al., 1999] Albert, R., Jeong, H., and Barabási, A. L. (1999). Diameter of the World Wide Web. *Nature*, 401.

[Almendral and Díaz-Guilera, 2007] Almendral, J. and Díaz-Guilera, A. (2007). Dynamical and spectral properties of complex networks. *New Journal of Physics*, 9:187.

[Andersen et al., 2002] Andersen, D., Feamster, N., Bauer, S., and Balakrishnan, H. (2002). Topology inference from BGP routing dynamics. In *Proc. of the 2nd ACM SIGCOMM Wkshp. on Internet measurment*, pages 243–248. ACM.

[Banerjee and Jost, 2009] Banerjee, A. and Jost, J. (2009). Spectral characterization of network structures and dynamics. *Dynamics On and Of Complex Networks*, pages 117–132.

[Barabási and Albert, 1999] Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509.

[Barabási et al., 2002] Barabási, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3–4):590–614.

[Barrat et al., 2004] Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proc. of the National Academy of Sciences of the United States of America*, 101(11):3747.

[Barthelemy et al., 2005] Barthelemy, M., Barrat, A., Pastor-Satorras, R., and Vespignani, A. (2005). Characterization and modeling of weighted networks. *Physica A: Statistical Mechanics and its Applications*, 346(1-2):34–43.

[Beyene et al., 2008] Beyene, Y., Faloutsos, M., Chau, D., and Faloutsos, C. (2008). The eBay graph: How do online auction users interact? In *Proc. of the IEEE Conf. on Computer Communications Wkshps.*, pages 1–6.

[Biggs, 1993] Biggs, N. (1993). *Algebraic graph theory*. Cambridge Univ Pr.

[Bilke and Peterson, 2001] Bilke, S. and Peterson, C. (2001). Topological properties of citation and metabolic networks. *Physical Review E*, 64(3):036106.

[Boccaletti et al., 2006] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308.

[Bollobás, 1998] Bollobás, B. (1998). *Modern graph theory*. Springer Verlag.

[Bollobás and Riordan, 2004] Bollobás, B. and Riordan, O. (2004). The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34.

[Bollobás et al., 2001] Bollobás, B., Riordan, O., Spencer, J., and Tusnády, G. (2001). The degree sequence of a scale-free random graph process. *Random Structures and Algorithms*, 18(3):279–290.

[Bonacich and Lloyd, 2001] Bonacich, P. and Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191–201.

[Bonato et al., 2009] Bonato, A., Hadi, N., Horn, P., Prałat, P., and Wang, C. (2009). A dynamic model for on-line social networks. *Algorithms and Models for the Web-Graph*, pages 127–142.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proc. of the seventh Intl. Conf. on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands. Elsevier Science Publishers B. V.

[Broder, 1986] Broder, A. (1986). How hard is it to marry at random? (On the approximation of the permanent). In *Proc. of the eighteenth annual ACM symposium on Theory of computing*, pages 50–58. ACM.

[Bunke, 2000] Bunke, H. (2000). Recent developments in graph matching. In *Proc. of the 15th Intl. Conf. on Pattern Recognition*, page 2117. Published by the IEEE Computer Society.

[Buriol et al., 2006] Buriol, L., Castillo, C., Donato, D., Leonardi, S., and Millozzi, S. (2006). Temporal analysis of the Wikigraph. In *Proc. Web. Intell. 2006*, pages 45–51.

[Burt, 2000] Burt, R. (2000). Decay functions. *Social Networks*, 22(1):1–28.

[Callaway et al., 2001] Callaway, D., Hopcroft, J., Kleinberg, J., Newman, M., and Strogatz, S. (2001). Are randomly grown graphs really random? *Physical Review E*, 64(4):41902.

[Chakrabarti and Faloutsos, 2006] Chakrabarti, D. and Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2.

[Chakrabarti et al., 2010] Chakrabarti, D., Faloutsos, C., and McGlohon, M. (2010). Graph mining: Laws and generators. *Managing and Mining Graph Data*, pages 69–123.

[Chen and Liu, 1997] Chen, S. and Liu, J. (1997). Statistical applications of the poisson-binomial and conditional bernoulli distributions. *Statistica Sinica*, 7:875–892.

[Chen et al., 2002] Chen, Y., Lim, K., Katz, R., and Overton, C. (2002). On the stability of network distance estimation. *ACM SIGMETRICS Performance Evaluation Review*, 30(2):21–30.

[Chung, 1997] Chung, F. (1997). Spectral graph theory (CBMS Regional Conf. Series in Mathematics, No. 92). *American Mathematical Society*, 3:8.

[Chung et al., 1994] Chung, F., Faber, V., and Manteuffel, T. (1994). An upper bound on the diameter of a graph from eigenvalues associated with its Laplacian. *SIAM J. Discrete Math.*, 7:443.

[Clauset et al., 2009] Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51:661.

[Costenbader and Valente, 2003] Costenbader, E. and Valente, T. (2003). The stability of centrality measures when networks are sampled. *Social networks*, 25(4):283–307.

[De Choudhury et al., 2010] De Choudhury, M., Mason, W. A., Hofman, J. M., and Watts, D. J. (2010). Inferring relevant social networks from interpersonal communication. In *Proc. of the 19th Intl. Conf. on World Wide Web*, WWW '10, pages 301–310, New York, NY, USA. ACM.

[Delvenne et al., 2010] Delvenne, J., Yaliraki, S., and Barahona, M. (2010). Stability of graph communities across time scales. *Proc. of the National Academy of Sciences*, 107(29):12755.

[Dhamdhere and Dovrolis, 2008] Dhamdhere, A. and Dovrolis, C. (2008). Ten years in the evolution of the internet ecosystem. In *Proc. of the 8th ACM SIGCOMM Conf. on Internet measurement*, pages 183–196. ACM New York, NY, USA.

[Diesner and Carley, 2005] Diesner, J. and Carley, K. M. (2005). Exploration of Communication Networks from the Enron Email Corpus. In *Proc. of the 2005 SIAM Wkshp. on Link Analysis, Counterterrorism and Security*, pages 3–14.

[Dong et al., 2009] Dong, Z., Wu, W., Ma, X., Xie, K., and Jin, F. (2009). Mining the structure and evolution of the airport network of China over the past twenty years. In *Proc. of the 5th Intl. Conf. on Advanced Data Mining and Applications*, page 115. Springer.

[Du et al., 2010] Du, N., Wang, H., and Faloutsos, C. (2010). Analysis of large multi-modal social networks: patterns and a generator. *Machine Learning and Knowledge Discovery in Databases*, pages 393–408.

[Eagle and Pentland, 2006] Eagle, N. and Pentland, A. (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268.

[Ebel et al., 2002] Ebel, H., Mielsch, L.-I., and Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical Review E*, 66(3):035103–+.

[Efraimidis and Spirakis, 2006] Efraimidis, P. S. and Spirakis, P. G. (2006). Weighted random sampling with a reservoir. *Information Processing Letters*, 97:181–185.

[Elmacioglu and Lee, 2005] Elmacioglu, E. and Lee, D. (2005). On six degrees of separation in DBLP-DB and more. *ACM SIGMOD Record*, 34(2):33–40.

[Erdös and Rényi, 1959] Erdös, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6(26):290–297.

[Eubank et al., 2004] Eubank, S., Kumar, V. S. A., Marathe, M. V., Srinivasan, A., and Wang, N. (2004). Structural and algorithmic aspects of massive social networks. In *Proc. of the 15th. annual ACM-SIAM symposium on discrete algorithms*, pages 718–727.

[Farkas et al., 2001] Farkas, I., Derenyi, I., Barabasi, A., and Vicsek, T. (2001). Spectra of "real-world" graphs: Beyond the semicircle law. *Physical Review E*, 64(2):26704.

[Goldstein et al., 2004] Goldstein, M., Morris, S., and Yen, G. (2004). Problems with fitting to the power-law distribution. *The European Physical Journal B*, 41(2):255–258.

[Good and Wang, 2005] Good, P. and Wang, R. (2005). *Permutation, parametric and bootstrap tests of hypotheses*. Springer New York.

[Hall et al., 2001] Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The NBER patent citations data file: Lessons, insights and methodological tools. NBER Working Papers 8494, National Bureau of Economic Research, Inc.

[Hanhijärvi et al., 2009] Hanhijärvi, S., Garriga, G. C., and Puolamäki, K. (2009). Randomization techniques for graphs. In *Proc. of the Ninth SIAM Intl. Conf. on Data Mining*, pages 780–791. SIAM.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

[Hill and Dunbar, 2003] Hill, R. and Dunbar, R. (2003). Social network size in humans. *Human Nature*, 14(1):53–72.

[Hoff et al., 2002] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

[Holme et al., 2004] Holme, P., Edling, C., and Liljeros, F. (2004). Structure and time evolution of an internet dating community. *Social Networks*, 26(2):155–174.

[Hu and Wang, 2009] Hu, H. and Wang, X. (2009). Evolution of a large online social network. *Physics Letters A*, 373:1105–1110.

[Huang et al., 2008] Huang, J., Zhuang, Z., Li, J., and Giles, C. L. (2008). Collaboration over time: characterizing and modeling network evolution. In *Proc. WSDM '08*, pages 107–116.

[Kleinberg, 2000] Kleinberg, J. M. (2000). Navigation in a small world. *Nature*, 406(6798):845.

[Kossinets, 2006] Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, 28(3):247–268.

[Kossinets and Watts, 2006] Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88–90.

[Krishnamurthy et al., 2008] Krishnamurthy, B., Gill, P., and Arlitt, M. (2008). A few chirps about twitter. In *Proc. of the first workshop on Online social networks*, WOSP '08, pages 19–24, New York, NY, USA. ACM.

[Kumar et al., 2006] Kumar, R., Novak, J., and Tomkins, A. (2006). Structure and evolution of online social networks. In *Proc. ACM SIGKDD '06*, pages 611–617.

[Kumar et al., 2000] Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tompkins, A., and Upfal, E. (2000). The Web as a graph. In *Proc. of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–10. ACM.

[Lahiri and Berger-Wolf, 2010] Lahiri, M. and Berger-Wolf, T. (2010). Periodic subgraph mining in dynamic networks. *Knowledge and Information Systems*, 24(3):467–497.

[Lahiri and Berger-Wolf, 2008] Lahiri, M. and Berger-Wolf, T. Y. (2008). Mining periodic behavior in dynamic social networks. In *Proc. of the IEEE Intl. Conf. on Data Mining*, pages 373–382.

[Langville et al., 2008] Langville, A., Meyer, C., and Fernández, P. (2008). Google's PageRank and beyond: the science of search engine rankings. *The Mathematical Intelligencer*, 30(1):68–69.

[Latapy and Magnien, 2006] Latapy, M. and Magnien, C. (2006). Measuring fundamental properties of real-world complex networks. *CoRR*, abs/cs/0609115.

[Latapy and Magnien, 2008] Latapy, M. and Magnien, C. (2008). Complex network measurements: Estimating the relevance of observed properties. In *Proc. IEEE INFOCOM '08*, pages 1660–1668.

[Latora and Marchiori, 2001] Latora, V. and Marchiori, M. (2001). Efficient behavior of small-world networks. *Phys. Rev. Lett.*, 87(19):198701.

[Leskovec and Faloutsos, 2006] Leskovec, J. and Faloutsos, C. (2006). Sampling from large graphs. In *Proc. of the 12th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, page 636. ACM.

[Leskovec et al., 2007] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Trans. on Knowledge Discovery from Data*, 1(1):2.

[Leskovec et al., 2005] Leskovec, J., Kleinberg, J. M., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proc. of the 11th ACM SIGKDD Intl. Conf. on Knowl. disc. and data mining*, pages 177–187.

[Ley, 2002] Ley, M. (2002). The DBLP computer science bibliography: Evolution, research issues, perspectives. In *SPIRE 2002: Proc. of the 9th Intl. Symposium on String Processing and Information Retrieval*, pages 1–10, London, UK. Springer-Verlag.

[Menezes et al., 2009] Menezes, G. V., Ziviani, N., Laender, A. H., and Almeida, V. (2009). A geographical analysis of knowledge production in computer science. *WWW 09: Proc. of the 18th Intl. Conf. on World Wide Web*, page 1041.

[Mislove et al., 2007] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proc. of the 7th ACM SIGCOMM Conf. on Internet measurement*, pages 29–42.

[Moody, 2004] Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2):213.

[Nanavati et al., 2006] Nanavati, A. A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S., and Joshi, A. (2006). On the structural properties of massive telecom call graphs: findings and implications. In *Proc. of the 15th ACM Intl. Conf. on Information and knowledge management*, pages 435–444, New York, NY, USA. ACM.

[Nascimento et al., 2003] Nascimento, M., Sander, J., and Pound, J. (2003). Analysis of SIGMOD's coauthorship graph. *ACM SIGMOD Record*, 32(3):8–10.

[Naumov et al., 2006] Naumov, V., Baumann, R., and Gross, T. (2006). An evaluation of inter-vehicle ad hoc networks based on realistic vehicular traces. In *MobiHoc '06: Proc. of the 7th ACM Intl. symposium on Mobile ad hoc networking and computing*, pages 108–119, New York, NY, USA. ACM.

[Nerur et al., 2005] Nerur, S., Sikora, R., Mangalaraj, G., and Balijepally, V. (2005). Assessing the relative influence of journals in a citation network. *Commun. ACM*, 48:71–74.

[Newman, 2001a] Newman, M. (2001a). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):16132.

[Newman, 2004] Newman, M. (2004). Analysis of weighted networks. *Physical Review E*, 70(5):56131.

[Newman, 2001b] Newman, M. E. J. (2001b). Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):25102.

[Newman, 2001c] Newman, M. E. J. (2001c). From the cover: The structure of scientific collaboration networks. *Proc. of the National Academy of Science*, 98:404–409.

[Newman, 2003] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

[Pallis et al., 2009] Pallis, G., Katsaros, D., Dikaiakos, M., Loulloudes, N., and Tassiulas, L. (2009). On the structure and evolution of vehicular networks. In *IEEE Intl. Symposium on Modeling, Analysis Simulation of Computer and Telecommunication Systems*, pages 1–10.

[Park et al., 2004] Park, S., Pennock, D., and Giles, C. (2004). Comparing static and dynamic measurements and models of the Internet's AS topology. In *Proc. IEEE INFOCOM*, volume 3, pages 1616–1627.

[Pedarsani et al., 2008] Pedarsani, P., Figueiredo, D. R., and Grossglauser, M. (2008). Densification arising from sampling fixed graphs. In *Proc. of the ACM SIGMETRICS Intl. Conf. on measurement and modeling of computer systems*, pages 205–216. ACM.

[Perra and Fortunato, 2008] Perra, N. and Fortunato, S. (2008). Spectral centrality measures in complex networks. *Physical Review E*, 78(3):36107.

[Raney et al., 2002] Raney, B., Voellmy, A., Cetin, N., Vrtic, M., and Nagel, K. (2002). Towards a microscopic traffic simulation of all of Switzerland. In *ICCS '02: Proc. of the Intl. Conf. on Computational Science-Part I*, pages 371–380, London, UK. Springer-Verlag.

[Ryan, 2008] Ryan, T. (2008). *Modern regression methods*. Wiley-Interscience.

[Sarkar and Moore, 2005] Sarkar, P. and Moore, A. W. (2005). Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, 7(2):31–40.

[Seber and Lee, 2003] Seber, G. A. and Lee, A. J. (2003). *Linear Regression Analysis*. Wiley-Interscience.

[Sharan and Neville, 2007] Sharan, U. and Neville, J. (2007). Exploiting time-varying relationships in statistical relational models. In *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 9–15. ACM.

[Shetty and Adibi, 2005] Shetty, J. and Adibi, J. (2005). Discovering important nodes through graph entropy the case of enron email database. In *LinkKDD '05: Proc. of the 3rd Intl. workshop on Link discovery*, pages 74–81, New York, NY, USA. ACM.

[Shi et al., 2007] Shi, X., Tseng, B., and Adamic, L. (2007). Looking at the blogosphere topology through different lenses. In *ICSWM 07: Proc. of the Intl. Conf. on Weblogs and Social Media*.

[Soffer and Vázquez, 2005] Soffer, S. N. and Vázquez, A. (2005). Network clustering coefficient without degree-correlation biases. *Phys. Rev. E*, 71(5):057101.

[Sulo et al., 2010] Sulo, R., Berger-Wolf, T., and Grossman, R. (2010). Meaningful selection of temporal resolution for dynamic networks. In *Proc. of the Eighth Wkshp. on Mining and Learning with Graphs*, MLG '10, pages 127–136, New York, NY, USA. ACM.

[Sun et al., 2007] Sun, J., Faloutsos, C., Papadimitriou, S., and Yu, P. S. (2007). GraphScope: parameter-free mining of large time-evolving graphs. In *Proc. of the 13th ACM SIGKDD Intl. Conf. on Knowl. disc. and data mining*, pages 687–696, New York, NY, USA. ACM.

[Sundaresan et al., 2007] Sundaresan, S. R., Fischhoff, I. R., Dushoff, J., and Rubenstein, D. I. (2007). Network metrics reveal differences in social organization between two fission–fusion species, Grevy's zebra and onager. *Oecologia*, 151(1):140–149.

[Thomas and Blitzstein, 2011] Thomas, A. C. and Blitzstein, J. K. (2011). Valued ties tell fewer lies: Why not to dichotomize network edges with thresholds. *ArXiv e-prints 1101.0788v2*.

[Vázquez et al., 2002] Vázquez, A., Pastor-Satorras, R., and Vespignani, A. (2002). Large-scale topological and dynamical properties of the Internet. *Physical Review E*, 65(6):66130.

[Wang, 1993] Wang, Y. (1993). On the number of successes in independent trials. *Statistica Sinica*, 3(2):295–312.

[Wang et al., 2003] Wang, Y., Chakrabarti, D., Wang, C., and Faloutsos, C. (2003). Epidemic spreading in real networks: An eigenvalue viewpoint. *IEEE Symposium on Reliable Distributed Systems*, 0:25.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

[West, 2001] West, D. B. (2001). *Introduction to graph theory*. Prentice Hall, NJ.

[Zhang and Horvath, 2005] Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128.