Revisiting Power-law Distributions in Spectra of Real World Networks

Nicole Eikmeier
Purdue University
Department of Mathematics
eikmeier@purdue.edu

David F. Gleich
Purdue University
Department of Computer Science
dgleich@purdue.edu

ABSTRACT

By studying a large number of real world graphs, we find empirical evidence that most real world graphs have a statistically significant power-law distribution with a cutoff in the singular values of the adjacency matrix and eigenvalues of the Laplacian matrix in addition to the commonly conjectured power-law in the degrees. Among these results, power-laws in the singular values appear more consistently than in the degree distribution. The exponents of the power-law distributions are much larger than previously observed. We find a surprising direct relationship between the power-law in the degree distribution and the power-law in the eigenvalues of the Laplacian that was theorized in simple models but is extremely accurate in practice. We investigate these findings in large networks by studying the cutoff value itself, which shows a scaling law for the number of elements involved in these power-laws. Using the scaling law enables us to compute only a subset of eigenvalues of large networks, up to tens of millions of vertices and billions of edges, where we find that those too show evidence of statistically significant power-laws.

CCS CONCEPTS

•**Networks** →**Network properties; Network structure;**

KEYWORDS

power-law distributions, real world networks, graph spectrum, degree distributions

1 INTRODUCTION AND MOTIVATION

Power-laws are a key component in any characterization of the networks gathered from the world wide web and other large information sources. These include web-crawls, online social networks, recommender systems, and many other examples [20, 29, 41]. There are quite a few places that power-laws may arise in the description of these networks. For instance, the degree distributions of these networks are often observed to have a power-law. Additionally, the eigenvalues of these networks are often observed to obey a power-law. Power-laws also arise in other types of structural statistics about the networks [48]. Properties of these power-laws are then used to generate realistic synthetic network models [2, 9] and to establish theory about why various algorithms work better than expected in networks of this type [16, 24, 32, 34, 58].

Towards these dual goals of building realistic models and generating useful theory, it is useful to have accurate information about the presence of power-laws in these real-world networks. The folk-lore about this is that networks have a power-law in their degree distribution with exponent between 2 and 3 (see, e.g., [12]). This finding does not always hold [50] and there is even contradictory evidence that these networks have power-law distributions [33, 51]. Moreover, there is a diverse literature on the implications of a power-law in the degree distribution for the behavior of the eigenvalues of the adjacency matrix and Laplacian matrix [11, 26, 43]. This literature argues that in specific models of a network, a power-law in the degree distribution implies a power-law in the eigenvalue distribution. It also hypothesizes that this may hold more broadly outside the specific model.

In this paper, we wish to revisit many of these empirical findings with the goal of providing new guidance on the presence of power-laws in three features of real-world networks:

(1) the degree distribution;
(2) the singular value distribution of the adjacency matrix;
(3) the eigenvalue distribution of the Laplacian matrix;

We built a large collection of real-world networks from the Stanford project, Facebook, and various other sources (See Section 4 and Appendix A for more about where our data originates.) We have computed the singular value distribution and Laplacian eigenvalue distribution exactly using a large cluster of high-performance computers [23]. To each distribution on each network, we fit the coefficients of a power-law distribution with cutoff in the tail following the methodology of Clauset [14]. (More specifically, we used the implementation by Nepusz [45], see the details our methods in Section 3). This fitting also included a test of significance, which allows us to gauge the reliability of the results. We call a power-law fit *significant* if it passes this test. This methodology resulted in the following observations.

(1) Many networks have a significant power-law in the tail of the degree distribution corresponding to the largest degree vertices, as well as the singular values of the adjacency matrix, and the eigenvalues of the Laplacian matrix. (Section 5.1)
(2) A significant power-law distribution is more likely to occur in the singular values of the adjacency matrix compared with the degree distribution. This means it is *more accurate* to assume a model where the singular values of the adjacency matrix have a power-law compared with the degree distribution. (Section 5.1)

KDD'17, August 13–17, 2017, Halifax, NS, Canada.

978-1-4503-4887-4/17/08…$15.00
DOI: http://dx.doi.org/10.1145/10.1145/3097983.3098128

(3) A significant power-law distribution in the degrees means there is a high probability for a significant power-law distribution in the singular values of the adjacency matrix and the eigenvalues of the Laplacian matrix. The converse does not hold. (Section 5.2)
(4) The coefficients of these power-laws vary from $2-10$ for all three distributions (degrees, adjacency singular values, and Laplacian eigenvalues). This is a much larger range than has been observed previously. (Section 6.1)
(5) The tail of the degree distribution and the Laplacian eigenvalues appear to behave identically and have essentially the same power-law distribution. That is, the power-law exponent and cutoff value are almost identical between the fitted distributions. (Section 6.1)
(6) The region of the distribution where the power-law fits appears to scale as $n^{2/3}$ for the degrees and Laplacian eigenvalues and between $n^{2/3}$ and $n^{1/2}$ for the singular values. (Section 6.3)
(7) We use observation 6 to test a number of large networks beyond those used to make observations 1-6 because it shows we would not have to compute the entire singular value and eigenvalue spectra. We find these observations hold on eight graphs up to 2 magnitudes larger than those used to form our hypotheses. (Section 7)

Overall, these findings refine our view of the power-laws and their relationships in real-world data of relevance to the community.

The presence of a power-law distribution in the singular values is an extremely powerful analytic property for understanding the real-world behavior of many types of matrix-based computations on large social networks (and why we would expect it to be far better than the worst case scenario). We believe the observation that the power-law in the singular values of the adjacency matrix is a more consistently observable feature than the power-law in the degree distribution to be a novel and useful outcome from this study. We plan to release all of the data and analysis codes necessary to replicate our findings once the paper is accepted.

We discuss additional implications of our results in the final section (Section 8). In comparison with past studies revisiting power-laws in networks [51], our focus is on the power-laws in the eigenvalues and singular values across a broad spectrum of networks. Regarding other conjectures and findings about the lack of power-laws in data [22, 42], we detail a few differences in our methodology in Section 3.4.

We provide the results of our power-law fits as well as our analytical tools in the github repository: https://github.com/eikmeier/powerlaw-spectra.

2 PRELIMINARIES

Our overall methodology is to fit power-law distributions to the degrees, singular values of the adjacency matrix, and the eigenvalues of the Laplacian matrix. We now concretely define these terms to clarify our specific usage. We present a summary in Table 1.

2.1 Power-laws

A set of values $x_1, \ldots, x_k$ satisfies a power-law if it is drawn from a probability distribution where $p(x) \propto x^{-\alpha}$ for some α. A set of values satisfies a power-law with a cutoff if $p(x) \propto x^{-\alpha}$ for all $x \geq x_{\min}$. The second case can be considered to describe a distribution with a power-law tail. Power-laws appear as *linear* relationships on a *log-log* plot because of the equivalent formulation:

$$\log p(y) = -\alpha \log y + c$$

Traditional methods to test for a power-law fit take advantage of this relationship and find a linear fit in the *log-log* plot. Unfortunately this type of graphical method is subject to errors [27]. We elaborate on a better methodology below due to Clauset et al. [14].

Table 1: Notation in our paper

$\boldsymbol{A}$	the adjacency matrix
$\boldsymbol{D}$	the degree matrix
$\boldsymbol{L}$	the Laplacian matrix
d_i	the degree of vertex i
$d_{\max}$	the largest degree
n	the number of vertices
m	the number of edges
p	the goodness-of-fit results for a power-law
α	the exponent of the power-law fit
$x_{\min}$	the cutoff for values in the power-law fit

2.2 Graphs and matrices

We start by defining some notation and basic concepts that we will use to describe our methods. Let $G = (V, E)$ be an unweighted, undirected graph without any loops. (All the graphs we work with are undirected.) Let $|V| = n$ be the number of vertices and $|E| = m$ the number of edges of G. The adjacency matrix of G is the symmetric matrix $\boldsymbol{A}$, where entry $A_{ij} = A_{ji}$ is equal to 1 if there is an edge between vertices i and j, and 0 otherwise. The degree of a vertex i, d_i, is equal to the number of vertices which have an edge connecting to vertex i. Let $d_{\max}$ be the largest degree. The Laplacian of a graph is $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ where $\boldsymbol{D}$ is the degree matrix, which is a diagonal matrix with $\boldsymbol{D}_{ii} = d_i$.

2.3 Eigenvalues and singular values

We briefly review a few facts to contextualize our methods. These topics can be studied in [56] for example. Any real-valued symmetric matrix has a set of n eigenvalues and an orthogonal set of n eigenvectors. The eigenvalues of the adjacency matrix range between $-d_{\max}$ and $d_{\max}$; the eigenvalues of the Laplacian range from 0 to $2d_{\max}$. For a real-valued symmetric matrix, the singular values are the absolute values of the eigenvalues, so for an adjacency matrix, they range from 0 to $d_{\max}$. Note that power-laws are not usually described for a mixture of positive and negative values. For this reason, we look at fitting power-law distributions to the *singular values* of the adjacency matrix and the eigenvalues of the Laplacian matrix.

3 THE DETAILS OF OUR FITTING METHODOLOGY

Recall that we are interested in a comparison between power-laws in the degree distribution, singular values of the adjacency matrix,

and eigenvalues of the Laplacian. For each distribution, we seek to estimate the power-law coefficient α and the cutoff value $x_{\min}$ – as well as a measure of the significance that we will discuss shortly. To simplify the setting, we consider only undirected, connected graphs without any self-loops. Thus, for any network with directed edges, we remove the directionality of the relationships, and extract the largest connected component.

3.1 Computing degrees, singular values, and eigenvalues

For each resulting undirected graph, we compute the degree distribution, all of the eigenvalues of the adjacency matrix (and by taking absolute values, all the the singular values as well), and all of the eigenvalues of the Laplacian matrix. The degree distribution is straightforward. To compute these eigenvalues, we used the MRRR algorithm as implemented in ScaLAPACK [18, 57], and executed an eigenvalue computation using a cluster of high performance computers at Sandia National Labs where we could load the entire matrix as a dense matrix and execute the $O(n^3)$ algorithm to find them. We are currently in the process of describing these computations in more detail [23].

At this point, we have three collections of non-negative values: the degrees, the singular values of the adjacency matrix, and the eigenvalues of the Laplacian matrix. We remove small elements of the singular values and eigenvalues because the power-law distributions stated above cannot model values of 0. More specifically, due to floating point approximation, we remove any value that is smaller than $2^{-52}n$.

3.2 Fitting power-law parameters

To fit the power-law parameters, α and $x_{\min}$, we use the maximum-likelihood algorithm developed by Clauset, Shalizi, and Newman [14]. Using this method is more accurate than the traditional method of fitting the slope of the *log-log* plot. More specifically, we use the implementation by Nepusz [45] that uses the BFGS algorithm to estimate the parameters. Additionally, this method and software calculates a goodness-of-fit parameter p that indicates whether the power-law fit is likely to be significant. This score is based on a randomized procedure. If the value $p > 0.1$, then this is evidence that the presence of a power-law is justified. We adopt the term *significant* to describe power-laws that pass this threshold.

3.3 Exceptions to our methodology

We note that this procedure worked for the vast majority of networks we mention in the next section. All told, we ran these procedures successfully for over 5000 distributions. We were able to fit the coefficients of the power-law for every single distribution. However, the goodness-of-fit computation reliably failed for three degree distributions (all synthetic networks); thus, we discarded these results as we cannot be confident in their significance. This is due to an issue of numerical precision in the software. In any case, we still have an extremely large database of results to study.

3.4 Critique

There are two weaknesses with this study that slightly temper our conclusions and we wish to address them. First, our observations 1-6 originate with data up to size 300k vertices, beyond which point it became computationally difficult to compute entire spectra of the networks. These networks originate from a variety of sources and include crawled as well as sampled networks. Recently, there have been studies on potential biases in power-law observations in networks of crawled data [1, 22, 42], which particularly apply to smaller networks. There are instances where the sampling procedure applied to the network causes properties to emerge that are not present in the underlying network [13, 52]. We agree our methodology cannot distinguish if the power-law originates due to the network collection methodology or reflects an underlying phenomenon. Although we note that just because there can be biases with crawling networks doesn't mean there will be problems.

To address these limitations, and towards the goal of studying larger networks, we include experiments on a set of large networks in Section 7 to investigate what happens for data two orders of magnitude larger than what we used to generate our hypotheses. These experiments support our observations. Furthermore, two of our large networks, *cit-Patents* and *wikipedia* are generated from a network data-dump, rather than a crawl.

Second, we wanted to study relationships between these power-laws, which meant we only used undirected graphs (and removed direction of edges in directed graphs), and we only considered the largest connected component. For this reason, existing negative results may not be directly comparable to ours. Meusel et al. for example, find that the degree distribution of a 3.56 Billion node web graph does not fit a power-law [42]. Nevertheless, our main interest is not in terms of power-laws in the degrees, but power-laws in the singular values and eigenvalues. Independent of the results with degree distribution power-laws, the observations about power-laws in singular values appear to be more robust than within the degrees – which has the potential to better inform future theoretical models of these networks.

4 DATA SETS AND MODELS

In this section, we present an overview on the data we use in our study. More detail on the source of each dataset is provided in Appendix A. These are all public datasets collected from various sources including the Pajek software [6], the SNAP collection [36], and the University of Florida collection [17].

Table 2 shows a quick view on all of our datasets. We have divided them into a number of groups based on common types of data. For some types of networks, we have a large number of samples (Facebook, Erdős, AS, Oregon, P2P), which we expect to be more highly related than the more general categories, and so these become their own categories. We also investigate four network models: graphs with a prescribed power-law degree distribution, graphs sampled from the copying model of graph evolution, graphs sampled from the preferential attachment model, and graphs sampled from the forest fire process.

4.1 Real-world data

At a high level, we break our real-world datasets into three categories: real-world data where a power-law might be a possibility (such as in collaboration networks, biology networks, citation networks, etc.); a subset of graphs where we do not expect power-law

Table 2: The types of networks we use in our studies, along with a rough order of magnitude of the sizes in vertices.

Type	Description	Sizes
Collab.	Co-authorship or collaboration networks defined by co-occurrence in author lists	100-100k
Biology	Protein-protein interaction networks	100-10k
Citation	Citations or references between a set of papers or other objections	1k-230k
Fiction	Networks drawn from fictional works	100-20k
Relational	A catch-all category for non-specific relational links including recommender system similarities, sports teams, trust networks, and others	100-20k
Social	Networks that model social interactions	100-100k
Tech.	Edges represent physical infrastructure including routers or power grid	5k-200k
Web	Hyperlink networks	1k-300k
Word	Various types of associations between words	100-100k
Low-dim	Networks with low-dimensional geometry (which should not have power-laws)	100-100k
Facebook	The Facebook 100 collection of networks	1k-50k
Erdős	9 collaboration networks centered on Erdős	100-5k
AS	(Autonomous systems) A large set of autonomous systems networks	100-25k
Oregon	Another set of AS networks	10k-11k
P2P	(Peer to peer) networks from Gnutella	1k-100k
RPL	(Random power-law) Random networks generated with a prescribed power-law degree distribution	13k
Copying	Networks from the copying model of graph evolution	1k-100k
PA	(Preferential attachment) networks	1k-10k
Forest fire	Networks generated from a forest-fire process	1k-100k

fits as the data comes from a low-dimensional space (low-dim). These include road networks and meshes. The third group is a number of networks that are more similar (as previously mentioned).

We provide a bit of detail on this third group of networks here. The AS type is autonomous systems network of routers on the internet, with edges as communications between two vertices [37]. The Oregon graphs are also autonomous systems [36]. Each of the Facebook networks are social networks where nodes represent people, and edges are a "friendship" between two nodes [55]. The P2P graphs are peer-to-peer networks from Gnutella, where nodes are agents and edges are again communication between two nodes [38]. Erdős is a collection of networks of Erdős's co-authors [6] collected over a few years.

4.2 Models

We now describe some relevant details about the models as there are often a variety of construction details that can vary, and we wish to be precise about our methods. Each model has a number of parameters. We picked parameters to explore a diversity of graphs generated by each model. We did not find any characteristic behavior in terms of the parameters and so we defer that information to the online data release.

In the copying model, we start with an initial clique graph and add vertices with the following process. A vertex arrives and picks a parent vertex uniformly at random. This new vertex then copies connections from the parent, but makes mistakes with probability α. A mistake drops a possible link. The graph is always undirected, so nodes can acquire new links via the copying process.

The forest fire model is similar [38]. We start with an initial clique. A vertex arrives and picks a parent uniformly at random. This new vertex then explores the local neighborhood of its parent via a forest-fire process that is akin to a randomly truncated breadth-first search. This process explores each node in the search frontier with probability q. The new node generates edges to any node that is explored in the process.

The preferential attachment model is the standard model [4] where new nodes connect to k-nodes chosen with probability proportional to their degree. The random power-law models generate a power-law distribution and then sample a graph using the Bayati-Saberi-Kim routine [7] with this degree distribution (or a slight perturbation necessary to ensure a graphical sequence).

5 PRESENCE OF POWER-LAWS

In this section we present the results of the fitting method on our data only in terms of whether or not the distributions support a power-law hypothesis via the goodness-of-fit test. In subsequent section, we will study these power-laws in more detail. This section serves to support the first three findings we reported on in the introduction.

5.1 Many classes of networks have power-laws

First, many networks have a significant power-law in the tail of the degree distribution corresponding to the largest degree vertices, as well as the singular values of the adjacency matrix and the eigenvalues of the Laplacian matrix. Table 3 lists the types of real-world networks and models that were described in the last section. For each type of network, we list the total number of networks and how many of each have a statistically significant power-law fit in the tail of the degree distribution, the singular values of the adjacency matrix, and eigenvalues of the Laplacian. We also list how many networks have a power-law distribution in two or more of those sets.

For example, from Table 3 we can see that of the 18 networks in category Oregon, 15 (83%) were found to have a power-law distribution in the degrees. Out of those with a power-law distribution in the degrees, 13 also have a power-law distribution in the eigenvalues of the adjacency matrix, which amounts to 72% of the total 18 networks.

As expected, the *low-dimensional networks* do not have power-law fits. Other classes with only a few significant power-law fits in the degrees include: fiction networks, P2P networks and Erdős's collaboration networks.

Note that our methodology is not perfectly sensitive as we only identify 85% of the networks with planted power-law distributions from the RPL experiments.

Note also that networks are far more likely to have significant power-laws in the singular values of the adjacency matrix than the degrees. Striking examples of this include the P2P and Erdős classes. Exceptions to this include forest fire networks, and random power-law networks, where this is almost true. This supports our point that it is more consistently true that real-world networks have a power-law in their adjacency singular values compared with

Table 3: For each type of real-world network or graph model, we list the total number of networks, and the number which have significant power-law fits in the degrees, adjacency singular value, Laplacian eigenvalues, and combinations.

		Distribution						Combinations						All	
	Number of Graphs	Degrees		Adjacency Sing. vals		Laplacian Eig. vals		Degrees & Adjacency		Degrees & Laplacian		Adjacency & Laplacian			
Biology	6	4	67%	6	100%	5	83%	4	67%	4	67%	5	83%	4	67%
Citation	6	4	67%	6	100%	5	83%	4	67%	4	67%	5	83%	4	67%
Collab.	13	5	38%	8	62%	5	38%	3	23%	4	31%	3	23%	2	15%
Fiction	3	1	33%	2	67%	0	0%	0	0%	0	0%	0	0%	0	0%
Relational	6	3	50%	3	50%	4	67%	2	33%	3	50%	2	33%	2	33%
Social	9	7	78%	8	89%	6	67%	6	67%	5	56%	5	56%	4	44%
Tech.	4	3	75%	4	100%	2	50%	3	75%	1	25%	2	50%	1	25%
Web	5	4	80%	4	80%	4	80%	3	60%	4	80%	3	60%	3	60%
Word	10	5	50%	9	90%	7	70%	5	50%	4	40%	6	60%	4	40%
Low-dim	2	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Facebook	100	74	74%	75	75%	80	80%	56	56%	73	73%	62	62%	56	56%
AS	855	831	97%	847	99%	846	99%	823	96%	824	96%	838	98%	816	95%
P2P	9	1	11%	6	67%	2	22%	1	11%	1	11%	2	22%	1	11%
Erdős	7	1	14%	7	100%	1	14%	1	14%	0	0%	1	14%	0	0%
Oregon	18	15	83%	16	89%	12	67%	13	72%	11	61%	11	61%	10	56%
Copying	163	69	42%	120	74%	81	50%	50	31%	60	37%	60	37%	43	26%
Forest fire	188	130	69%	82	44%	135	72%	59	31%	108	57%	60	32%	49	26%
PA	81	66	81%	68	84%	81	100%	53	65%	66	81%	68	84%	53	65%
RPL	20	17	85%	16	80%	18	90%	15	75%	17	85%	16	80%	15	75%

their degrees. This observation is much weaker for the Laplacian eigenvalues, for a reason that we will discuss shortly when we analyze the power-law fits themselves (Section 6.1).

5.2 Relationships between power-laws

In our second analysis, we study combinations of power-laws. We only do this for classes of networks where at least 40% of networks had *all three* power-laws to avoid drawing conclusions from small sample sizes. This removes the classes: Collaboration, Fiction, Technological, Relational, P2P, and Erdős. We still include the network models for reference.

We study these relationships in terms of conditional probabilities. Consider the probability that, given a power-law fit in the degrees, there is a power-law fit in the singular values of the adjacency matrix; denote this as $\mathrm{P}[\boldsymbol{A}|\boldsymbol{D}]$. In contrast consider the likelihood that given a power-law fit in the singular values of the adjacency matrix, there is a power-law fit in the degrees. Denote this as $\mathrm{P}[\boldsymbol{D}|\boldsymbol{A}]$. We use a similar notation regarding the Laplacian eigenvalues.

We list the probabilities in Table 4. Observe that $\mathrm{P}[\boldsymbol{A}|\boldsymbol{D}]$ is almost always larger than $\mathrm{P}[\boldsymbol{D}|\boldsymbol{A}]$ which is to say that a power-law fit in the tail of the degrees gives a high likelihood for a power-law fit in the tail of the singular values, but not vice-versa. Similarly $\mathrm{P}[\boldsymbol{L}|\boldsymbol{D}]$ is usually larger than $\mathrm{P}[\boldsymbol{D}|\boldsymbol{L}]$, which means that a power-law fit in the tail of the degrees likely implies a power-law fit in the tail of the Laplacian eigenvalues. Both of these relationships have been studied in a variety of theoretical settings in graph models such as the Chung-Lu graphs [12, 19, 43]. Given the diversity of real-world data explored here, it is reassuring to see that these theoretical predictions have meaningful real-world evidence.

Table 4: Conditional probabilities that a power-law distribution in one feature gives a power-law distribution in another; *D* stands for degrees, *A* stands for the singular values of the adjacency matrix, and *L* stands for eigenvalues of the Laplacian. The first column for example is the probability that there is a significant power-law distribution in the singular values of the adjacency matrix given that there is a significant power-law distribution in the degrees. For the first group of measurements, we combine the data and compute probabilities in the summary of *Other* class.

Type	$\mathrm{P}[\boldsymbol{A}\|\boldsymbol{D}]$	$\mathrm{P}[\boldsymbol{D}\|\boldsymbol{A}]$	$\mathrm{P}[\boldsymbol{L}\|\boldsymbol{D}]$	$\mathrm{P}[\boldsymbol{D}\|\boldsymbol{L}]$	$\mathrm{P}[\boldsymbol{L}\|\boldsymbol{A}]$	$\mathrm{P}[\boldsymbol{A}\|\boldsymbol{L}]$
Biology	1.0	0.67	1.0	0.8	0.83	1.0
Citation	1.0	0.67	1.0	0.8	0.83	1.0
Social	0.86	0.75	0.71	0.83	0.62	0.83
Web	0.75	0.75	1.0	1.0	0.75	0.75
Word	1.0	0.56	0.8	0.57	0.67	0.86
Sum.* of *Other	0.92	0.67	0.88	0.78	0.73	0.89
Facebook	0.76	0.75	0.99	0.91	0.83	0.78
AS	0.99	0.97	0.99	0.97	0.99	0.99
Oregon	0.87	0.81	0.73	0.92	0.69	0.92
Copying	0.72	0.42	0.87	0.74	0.5	0.74
Forest fire	0.45	0.72	0.83	0.8	0.73	0.44
PA	0.8	0.78	1.0	0.81	1.0	0.84
RPL	0.88	0.94	1.0	0.94	1.0	0.89

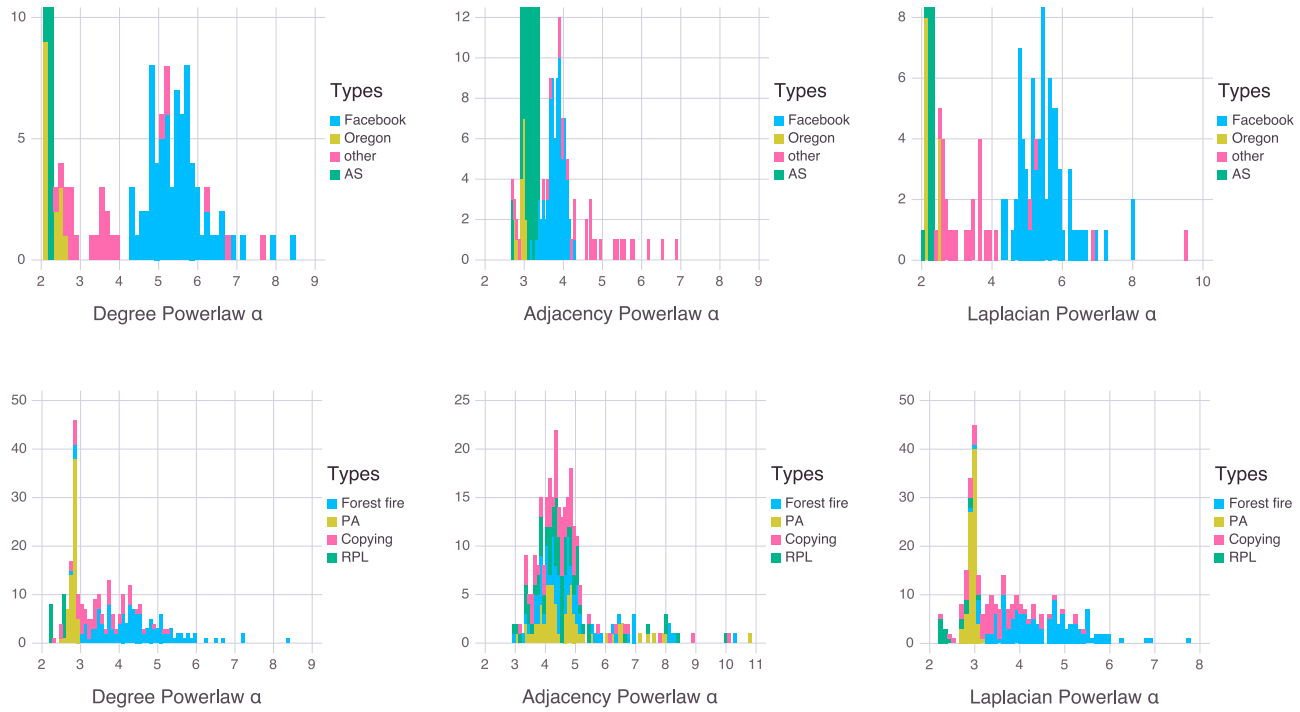

Figure 1: Exponents of statistically significant power-law distributions. In the first column is the exponent for the degree distribution, in the middle column the exponent in singular values of the adjacency matrix, and in the third column the exponent for the Laplacian eigenvalues. The top row is of the real networks, while the bottom row is of the graph models. The class "other" includes all small classes of real-world networks where power-laws are common (see the text).

6 ANALYSIS OF THE POWER-LAWS

In this section we consider the exponents of the power-law distributions, and we elaborate on several observations (points 4-6) from the introduction. For some of these studies, we wish to draw conclusions over multiple types of networks. For this reason, we create a new class of network called "Other" that consists of the classes Biology, Citation, Social, Web, and Word. (These are all classes where at least 40% of networks had a significant power-law in all three distributions. The findings are robust to nearby choices for this 40% threshold and the goal is to exclude classes of networks that seem to reliably *lack* power-laws.)

6.1 Characteristics of the power-laws

The first characteristic of the power-law fits we examine are the exponents α. In Figure 1 we plot the exponents of power-law distributions in the degrees, singular values of the adjacency matrix, and eigenvalues of the Laplacian matrix. We consider both real world networks and models. We see that the majority of exponents of the power-law distributions vary from $2-10$, and notice particularly that they are often greater than 3, which is much larger than observed previously (e.g., [12, 50]).

Next, we notice that the exponent for power-law of Laplacian eigenvalues is often nearly identical to the exponent of the power-law of degree. This has been conjectured for a variety of models [19]. Figure 2 shows the relationship between these power-law fits, which almost perfectly fits to the line $\alpha_{\text{Laplacian}} = \alpha_{\text{degree}}$, matching the theory well outside of its regime where it should apply. Furthermore, the values for the cutoff value ($x_{\min}$) in the degrees and Laplacian eigenvalues are nearly identical. This is to say that not only do the power-law fits have the same exponent, but they also fit to the same range of values. Thus, the tail of the degree distribution and the Laplacian eigenvalues appear to have essentially the same power-law distribution.

Finally, we study the conjecture that the power-law in the adjacency singular values should be $2\alpha_{\text{degree}} - 1$ when there is a power-law in both [11]. Figure 3 shows the relationships between these exponents. We see no hints of this scaling law in the real-world data. But, both the random power law graphs and the forest fire graphs show some agreement with this scaling. Thus, whereas the Laplacian result appears accurate, the adjacency result is not.

6.2 Consistency across network samples

Another observation we make is that the power-law distribution in the singular values of the adjacency matrix often appears to be more consistent when given multiple samples of the same network.

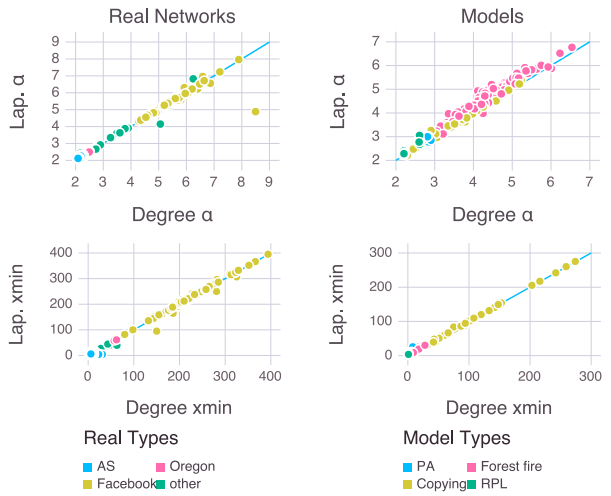

Figure 2: Networks with statistically significant power-law distributions in their degrees and eigenvalues of the Laplacian matrix. Top row: the exponent of the power-law fit in the degrees vs. the exponent of the power-law fit in the Laplacian eigenvalues. Bottom row: the cutoff of the power-law fit in the degrees vs. the power-law fit in the Laplacian eigenvalues. On the left are the real world networks, and on the right are the models. We plot the line $y = x$ on the same axes. A few networks are outliers from the line: Newman's *netscience*, biology (protein-protein) network *dmela*, and social network *Caltech*. We could not find any common features of these outliers.

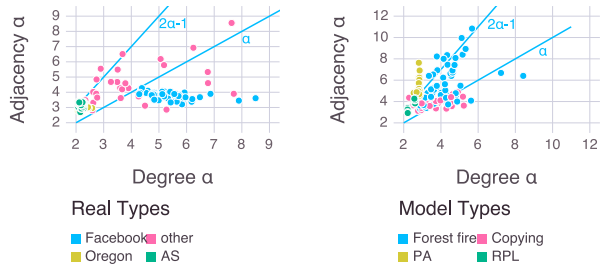

Figure 3: The exponent of the power-law fit in the degrees versus the exponent of the power-law fit in the singular values. We also plot the line $\alpha_{\text{Adj.}} = 2\alpha_{\text{Deg.}} - 1$, for comparison with results in the literature. The class "other" includes all small classes of real-world networks where power-laws are common (see the text).

We illustrate this via studying the density of the exponents over instances with multiple similar types of networks: Facebook, AS, and Oregon (Figure 4). We include the models Copying and PA (preferential attachment) as well. In the Facebook networks, for instance, the degree power-law exponents vary considerably, whereas the singular value power-law has a sharp distribution about 4. For Oregon and Copying, we see similar behavior. For AS networks, the singular values may have a slightly larger region, the preferential attachment networks are a counter-example.

6.3 Behavior of the cut-off

Our final observation about the nature of these power-laws reflects the number of entries where the power-law applies. Recall that $x_{\min}$ is the smallest value contained in the power-law distribution. Given a cutoff value $x_{\min}$, we compute the size of the tail of the distribution, i.e., the number of values larger than $x_{\min}$. We show the size of the tail relates to the size of the network in Figure 5 for both the degree and singular value power-laws. (The Laplacian power-law will behave almost identically to the degree power-law based on the discussion in Section 6.1).

The size of the tail appears to scale as $n^{2/3}$ in the degree distribution. A least-squares fit produces essentially the same result ($n^{0.67}$). For the singular values tails, the Facebook class shows the same $n^{2/3}$ scaling; but the other networks show scaling closer to $n^{1/2}$. (The least-squares fit chooses $n^{0.51}$). Both of these scenarios ($n^{2/3}, n^{1/2}$) indicate a shrinking fraction of the network where the power-law applies as the network size increases. However, they also provide useful practical advice about the region where "large degrees" and "large singular values" lie – which is important to understand for analyzing algorithms on these networks as well as designing models.

7 LARGE NETWORKS

The networks discussed up to this point have been relatively small, topping out at around $300k$ vertices. There is a computational hurdle in computing entire eigenvalue and singular value spectra for graphs with a million or more vertices in that most approaches need $O(n^2)$ memory. Observation 6, however, offers an approach: as discussed in Section 6.3, a lower bound on the number of degrees or singular values we expect to be included in a power-law tail is $n^{(1/2)}$. This suggests we need not compute all eigenvalues when testing for a power-law distribution.

We considered 8 large graphs from SNAP [36], MPI [44], and Wikipedia, listed in Table 5 along with information about the power-laws in the data. The *friendster, orkut, youtube, flickr, livejournal* data are all social networks, *skitter* is a technological router graph, *patents* a citation network, *wikipedia* a web network formed by Wikipedia articles and their categories where an edge occurs when there is a link between a pair of articles. Note that *wikipedia*, and *patents* are created from database dumps rather than crawls, whereas *orkut, youtube, flickr, livejournal* are all crawled. These results support our findings from previous observations: the singular values of the adjacency matrix are more likely to have a power-law than the degrees. With regards to the cutoff, the vast majority of values are included in these power-law, with exceptions noted below. Thus, these results show that hypotheses formed from graphs up to 300k vertices are also supported on data 100 times larger.

Details of experiment. We chose to restrict our analysis of the large networks to the top $n^{(1/2)}$ degrees and top $n^{(1/3)}$ singular values due to the computational complexity of testing more, and on the assumption that both of these regions would contain power-laws if they are present. The top $n^{(1/3)}$ singular-values of the adjacency matrix of each network were computed using ARPACK [35] (set

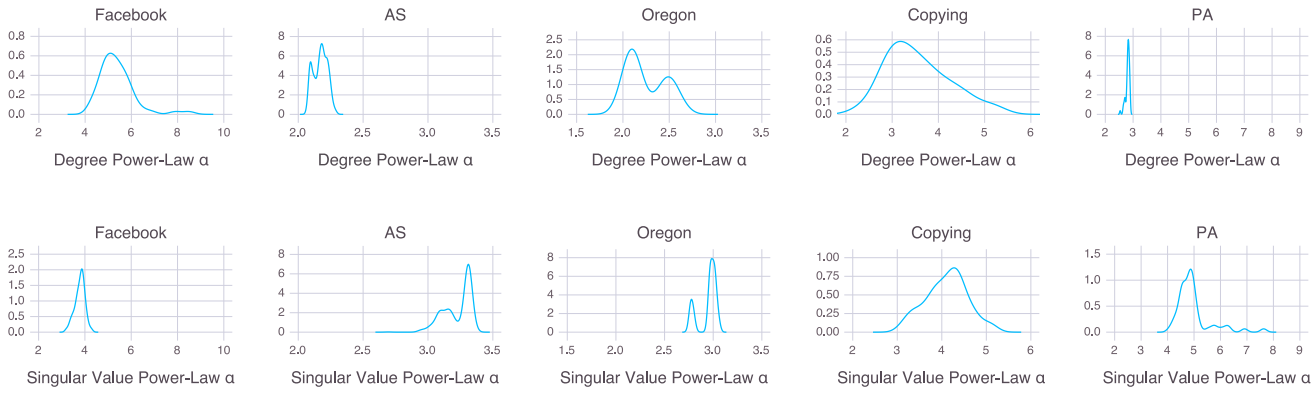

Figure 4: Density estimates of the power-law exponent α for five classes: Facebook, AS, Oregon, Copying, and PA. On the top are the exponents of the power-law fit in the degree distribution, and on the bottom are the power-law exponents of the singular values of the adjacency matrix. The singular value exponents are more consistent in 3 of the 5 types (Facebook, Oregon, Copying), slightly less consistent in one (AS), and more variable in one (PA).

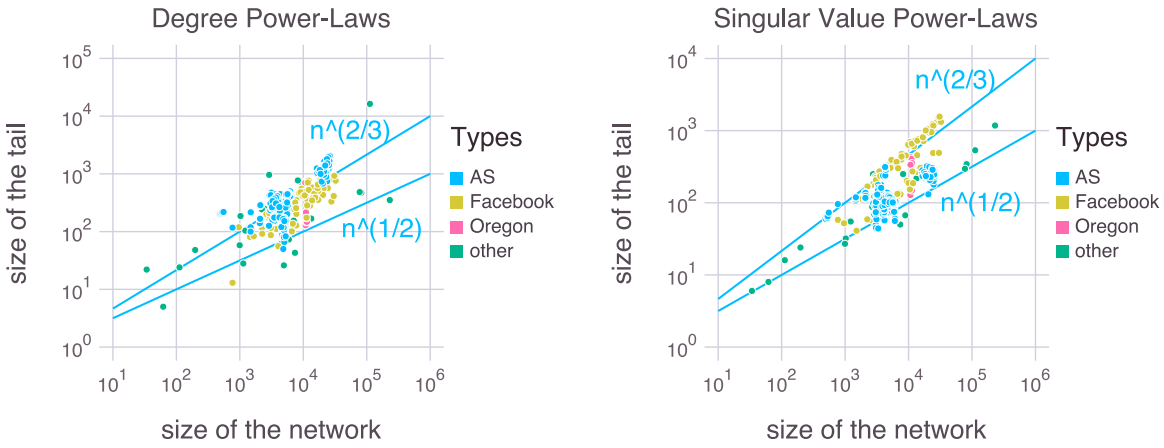

Figure 5: Log-Log plots of the size of the network versus the size of the tail in the power-law distributions, (the number of values greater than $x_{\min}$). The lines $n^{1/2}$ and $n^{2/3}$ are given on the same plot for reference. On the left are the degree distributions and on the right are the singular value distributions. The class "other" includes all small classes of real-world networks where power-laws are common (see the text).

to a tolerance of 10^{-8}). A power-law distribution was fit using the plfit method discussed in Section 3.2. In most cases, the cutoff (see section 6.3) included all or nearly all of the data points which we included for testing. Exceptions are the degree distribution of *livejournal* and the singular value distribution of *flickr*, which only included about half of the values. There is one network, *friendster*, for which the plfit software fails on the degrees when fitting to a discrete distribution (see Section 3.3 for a short discussion on this). We instead fit *friendster* to a continuous distribution. This gives similar parameters to the discrete fitting procedure in the other cases we evaluated. The *friendster* network is found to have a significant power-law distribution, but there are only 45 values greater than $x_{\min}$.

8 DISCUSSION AND IMPLICATIONS

A power-law in the degree distribution of a network was one of the hallmarks of the early study on web graphs and other types of information networks [4]. This initial focus on power-laws then led to a number of theoretical studies about the presence of power-laws in other features of the network including the singular value distribution and Laplacian eigenvalues [11, 19, 43] – as well as work critical of this finding [1]. We have conducted a wide-ranging evaluation of these conjectured relationships and discovered that: (i) the presence of a power-law in the largest singular values is more reliable than power-laws in the degree distribution; (ii) this power-law often applies to at least $n^{(1/2)}$ largest singular values, and for some classes of networks, up to $n^{(2/3)}$. Moreover, we find

Table 5: Results of fitting a power-law distribution to large graphs. A power-law was found to be significant with exponent α if it is labeled α^*, and insignificant if labeled α^-. Entries labeled $\alpha^\diamond$ are special cases, and discussed in the text.

Graph	Nodes	Edges	Deg. PL	Adj. PL
youtube	1.13M	2.99M	**2.5***	**4.2***
flickr	1.62M	15.5M	**3.89***	**3.09***
skitter	1.69M	11.1M	2.25$^-$	**3.5***
orkut	3.1M	117M	**2.62***	**4.76***
patents	3.76M	16.5M	**4.18***	**5.19***
livejournal	5.2M	48.9M	**3.3***	**3.55***
wikipedia	9.47M	107M	**2.46***	**4.03***
friendster	65.6M	1.8B	10.5$^\diamond$	**5.04***

compelling empirical evidence of the relationships from [19], which posits that the Laplacian eigenvalues and the degree distribution should have similar power-law exponents and behavior. We have attempted to address limitations of our methodology by testing large networks drawn from complete databases as well (Section 7).

Understanding the structure of real world networks allows us to take advantage of that structure for faster and better computation. In particular, we suspect the results of the reliable power-law in the singular values to be a useful property for characterizing the extremely fast convergence of many matrix-based algorithms on these types of networks. Beyond matrix-based algorithms, there are a variety of situations where these networks do not elicit worst-case behavior—for example maximum cliques appear to be easy to find [49]—having insight into the spectra of matrices from these networks provides another possible avenue to explain these results. Finally, we found that empirical power-law exponents are far larger than previously discussed, which may impact how we create random networks for statistical tests on networks and the applicability of existing results on power-law graphs (e.g. [16, 24, 32, 34, 58]).

Acknowledgments We are grateful to C. Seshadri for discussing ideas with us. Eikmeier acknowledges the support of NSF STC award CCF-093937 for the Center for the Science of Information. Gleich is partially supported by NSF CAREER award CCF-1149756, NSF award IIS-1546488, NSF STC award CCF-093937, the DARPA SIMPLEX program, and the Sloan Foundation.

A SOURCES OF DATA

We used data from a number of publicly available sources, but primary among them are the SNAP repository [36], Pajek collection [6], the University of Florida sparse matrix collection [17], and the Facebook100 [55]. There is overlap and duplication of networks between these groups. We also used a number of smaller collections of networks. We have attempted to cite a large subset of the suggested sources for the networks we have used.

Small collections. *Fictional social networks* [3]; *Collaboration* [10]; *Relational (Dictionary) blondel2004-graph-similarity*; *Biology* [30, 53]; *Technological* [54]; *Web* [15]; *Low-dim. (Mesh)* [21].

Newman's collection [46, 47, 47] *lesmis* [31]; *dolphins* [40]

Arenas's collection: *Jazz* [25], *email* [28], *PGP* [8],

SNAP. We used the following networks from SNAP. **Collaboration** *ca-AstroPh, ca-CondMat, ca-GrQc, ca-HepPh, ca-HepTh* [38]; **Social** *email-Enron, soc-Epinions1, soc-Slashdot0811, soc-Slashdot0902, wiki-Vote* [39];**Web** *web-NotreDame* [5].

Pajek. We used the following networks from Pajek. **Citation** *Kohonen, Lederberg, patents_main, SciMet, SmaGri, Zewail*; **Collaboration** *geom*; **Relational** *CSPhd, EVA*; **Technological** *USpowerGrid* **Web** *California, EPA*; **Word** *dictionary28, EAT_RS, FA, foldoc, ODLIS, Reuters911, Roget, Wordnet3.*

REFERENCES

[1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. In *STOC fi05: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 694–703. ACM Press, 2005.

[2] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *STOCSTOC '00 Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 171–180. ACM, 2000.

[3] R. Alberich, J. Miro-Julia, and F. Rossello. Marvel universe looks almost like a real social network. *arXiv*, cond-mat.dis-nn:0202174, 2002.

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[5] A.-L. Barabási, H. Jeong, and R. Albert. The www network. Accessed online via http://www.nd.edu/~networks/resources.htm in May 2009, 1999.

[6] V. Batagelj and A. Mrvar. Pajek datasets. http://vlado.fmf.uni-lj.si/pub/networks/data/, 2006.

[7] M. Bayati, J. Kim, and A. Saberi. A sequential algorithm for generating random graphs. *Algorithmica*, 58(4):860–910, 2010. 10.1007/s00453-009-9340-1.

[8] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas. Models of social networks based on social distance attachment. *Phys. Rev. E*, 70(5):056122, Nov 2004.

[9] A. Bonato, J. Janssen, and P. Pra llat. Geometric protean graphs. *Internet Mathematics*, 8(1-2):2–28, 2012.

[10] F. Bonchi, P. Esfandiar, D. F. Gleich, C. Greif, and L. V. Lakshmanan. Fast matrix computations for pairwise and columnwise commute times and Katz scores. *Internet Mathematics*, 8(1-2):73–112, 2012.

[11] F. Chung, L. Lu, and V. Vu. Eigenvalues of random power law graphs. *Annals of Combinatorics*, 7:21–33, 2003.

[12] F. Chung, L. Lu, and V. Vu. Spectra of random graphs with given expected degrees. In *Proceedings of the National Academy of Sciences*, volume 100, pages 6313–6318. PNAS, May 2003.

[13] A. Clauset and C. Moore. Accuracy and scaling phenomena in internet mapping. *Physical Review Letters*, 94(1):018701, 2005.

[14] A. Clauset, C. R. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[15] P. G. Constantine and D. F. Gleich. Using polynomial chaos to compute the influence of multiple random surfers in the PageRank model. In *Proceedings of the 5th Workshop on Algorithms and Models for the Web Graph*, volume 4863 of *LNCS*, pages 82–95. Springer, 2007.

[16] C. Cooper, T. Radzik, and Y. Siantos. A fast algorithm to find all high degree vertices in power law graphs. In *Proceedings of the 21st International Conference on World Wide Web*, pages 1007–1016. ACM, 2012.

[17] T. A. Davis and Y. Hu. The University of Florida sparse matrix collection. *ACM Trans. Math. Softw.*, 38(1):1:1–1:25, December 2011.

[18] I. S. Dhillon, B. N. Parlett, and C. Vömel. The design and implementation of the MRRR algorithm. *ACM Trans. Math. Softw.*, 32(4):533–560, December 2006.

[19] R. Elsässer. Toward the eigenvalue power law. In *International Symposium on Mathematical Foundations of Computer Science*, pages 351–362. Springer, 2006.

[20] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, 1999.

[21] J. R. Gilbert and S.-H. Teng. MESHPART: Matlab mesh partitioning and graph separator toolbox. http://www.cerfacs.fr/algor/Softs/MESHPART/, February 2002.

[22] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in Facebook: A case study of unbiased sampling of OSNs. In *IEEE INFOCOM*. IEEE, 2010.

[23] D. F. Gleich. A repository of graph spectra. In preparation, 2017.

[24] D. F. Gleich and C. Seshadhri. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In *KDD2012*, pages 597–605, Aug. 2012.

[25] P. M. Gleiser and L. Danon. Community structure in jazz. *Advances in Complex Systems*, 06(04):565–573, 2003.

[26] K.-I. Goh, B. Kahng, and D. Kim. Spectra and eigenvectors of scale-free networks. *Physical Review E*, 64(5):051903, 2001.

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

[28] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68:065103, Dec 2003.

[29] B. A. Huberman. *The Laws of the Web.* The MIT Press, Cambridge, Massachusetts, 2001.

[30] G. Klau. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, 10(Suppl 1):S59, January 2009.

[31] D. E. Knuth. *The Stanford GraphBase: A Platform for Combinatorial Computing.* Addison-Wesley, 1993.

[32] V. Kurauskas and M. Bloznelis. Large cliques in sparse random intersection graphs. *arXiv*, math.CO:1302.4627, 2013.

[33] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[34] M. Latapy. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical Computer Science*, 407:458–473, 2008.

[35] R. B. Lehoucq and D. C. Sorensen. Deflation techniques for an implicitly restarted arnoldi iteration. *SIAM J. Matrix Anal. Appl.*, 17(4):789–821, 1996.

[36] J. Leskovec. The Stanford large network dataset collection. http://snap.stanford.edu/data/index.html, 2016.

[37] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 177–187, New York, NY, USA, 2005. ACM.

[38] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), Marsh 2007.

[39] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, September 2009.

[40] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: Can geographic isolation explain this unique trait? *Behavioral Ecology and Sociobiology*, 54(4):pp. 396–405, 2003.

[41] A. Medina, I. Matta, and J. Byers. On the origin of power laws in internet toplogies. *ACM SIGCOMM Computer Communication Review*, 30(2), 2000.

[42] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer. The graph structure in the web - analyzed on different aggregation levels. *The Journal of Web Science*, (1):33–47, 2015.

[43] M. Mihail and C. Papadimitriou. On the eigenvalue power law. In *RANDOM '02 Proceedings of the 6th International Workshop on Randomization and Approximation Techniques*, pages 254–262, 2002.

[44] A. Mislove, M. Marcon, K. P. gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet Measurement*, ICM '07, pages 29–42, New York, NY, October 2007. ACM.

[45] T. Nepusz. plfit software. https://github.com/ntamas/plfit, 2016.

[46] M. Newman. Network datasets. http://www-personal.umich.edu/~mejn/netdata/, 2006.

[47] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[48] E. Papalexakis, B. Hooi, K. Pelechrinis, and C. Faloutsos. Power-hop: A pervasive obervation for real complex networks. *PLoS ONE*, 11(3), 2016.

[49] R. A. Rossi, D. F. Gleich, and A. H. Gebremedhin. Parallel maximum clique algorithms with applications to network analysis. *SIAM Journal on Scientific Computing*, 37(5):C589–C616, 2015.

[50] A. Sala, L. Cao, C. Wilson, R. Zablit, H. Zheng, and B. Y. Zhao. Measurement-calibrated graph models for social network experiments. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 861–870, New York, NY, USA, 2010. ACM.

[51] A. Sala, H. Zheng, B. Y. Zhao, S. Gaito, and G. P. Rossi. Brief announcement: Revisiting the power-law degree distribution for social graph analysis. In *Proceedings of the 29th ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing*, PODC '10, pages 400–401. ACM, 2010.

[52] G. Schoenebeck. Potential networks, contagious communities, and understanding social network structure. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1123–1132. ACM, 2013.

[53] R. Singh, J. Xu, and B. Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *PNAS*, 105(35):12763–12768, 2008.

[54] C. (The Cooperative Association for Internet Data Analyais). Network datasets. http://www.caida.org/tools/measurement/skitter/router_topology/, 2005. Accessed in 2005.

[55] A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of facebook networks. *Physica A*, 391(16):4165–4180, 2012.

[56] P. Van Mieghem. *Graph spectra for complex networks.* Cambridge University Press, 2011.

[57] C. Vömel. Scalapack's MRRR algorithm. *ACM Trans. Math. Softw.*, 37(1):1:1–1:35, January 2010.

[58] D. J. Watts. *Six Degrees: The Sceince of a Connected Age.* W. W. Norton and Company, 500 Fifth Avenue, New York, N.Y. 10110, 2003.