THE WEB AS A GRAPH

Following Baldi [BFS03], Chapter 3

SATU VIRTANEN

satu@tcs.hut.fi

October 1 and 8, 2003

T-122.101 Modeling and mining the Web

OUTLINE $\frac{1}{3}$

OUTLINE $\frac{2}{3}$

OUTLINE $\frac{3}{3}$

1 Internet and Web Graphs

Both the Internet ($=$ routers, domains, cables $=$ a hardware-based physical network) and the World Wide Web ($=$ sites, pages, hyperlinks $=$ a content-based semantical network) have been intensively studied for the past decade or so.

Over the past few years, the study of complex systems as graphs has grown rapidly into nearly a popular-science topic, due to its interdisciplinary and widely applicable nature. The aforementioned networks have become prototypes and common objects of study for the complex systems community.

Network

A network consists of a set of *nodes* and usually several *connections* between the nodes. The nodes or the connections may contain some additional information, such as labels or weights.

Networks are common models of complex systems in many fields of science; there are numerous examples in engineering, biology, sociology, and even linguistics.

Graph

The formalism for discussing networks is a *graph*. A *graph* $G = (V, E)$ consists of two distinct sets: $V = \{v_1, \ldots, v_n\}$ contains the *vertices* of the graph, and E contains the *edges* of the graph. Each edge is a pair of vertices.

The *order* $|G|$ of the graph is the number of vertices, denoted by $|V|$ or briefly n. The number of edges is called the *size* of the graph, denoted by $|E|$ or m.

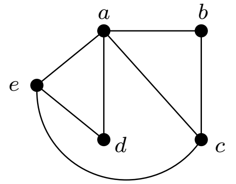
Density

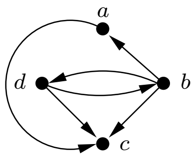
The *density* of a graph $G = (V, E)$ is defined as the ratio of edges present to the maximum possible

$$\delta = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)}$$

(for undirected graphs where the edges are unordered pairs).

EXAMPLES OF GRAPHS





An undirected graph $G_1 = (V_1, E_1)$ with $V_1 = \{a, b, c, d, e\}$ and $E_1 = \{(a, b), (a, c), (a, d), (a, e), (b, c), (c, e), (d, e)\}$, and a directed graph $G_2 = (V_2, E_2)$ with $V_2 = \{a, b, c, d\}$ and $E_2 = \{\langle a, c\rangle, \langle b, a\rangle, \langle b, c\rangle, \langle b, d\rangle, \langle d, b\rangle, \langle d, c\rangle\}$.

THE WEB AS A GRAPH

The vertices of the graph are, for example, either individual web pages or entire websites. The hyperlinks are the edges; their direction may be either ignored or incorporated in the graph.

The Web graph is very sparse, i.e., has small density. An average web page certainly does not link to but a miniscule fraction of the existing web pages.

The linking graph of web pages is very dynamic; it undergoes constant change as the numerous independent authors edit their websites. Also new nodes (i.e., pages or sites) appear practically all the time, while some of the existing ones are removed.

UNIFORM RANDOM GRAPHS

Gilbert's [Gil59] $G_{n,p}$ model for generating *random graphs* is the following: fix the order of the graph n and a probability p.

For each unordered pair of vertices (of which there are $\binom{n}{2}$), include the corresponding edge in the graph with probability p, otherwise leave it out. The resulting graph has expected density equal to p.

Each graph with m edges has probability $p^m(1-p)^{\binom{n}{2}-m}$ of being constructed; hence this model is referred to as the *binomial model*. A similar model $G_{n,m}$, where instead of p, the number of edges m is fixed, was studied by Erdős and Rényi [ER76a, ER76b].

NEIGHBORHOOD

The *neighborhood* of a vertex $v \in V$, denoted by $\Gamma(v) \subseteq V$, is the set of vertices that share an edge with v, $\Gamma(v) = \{u \mid (v, u) \in E\}$.

Note that a vertex itself is not considered to be a part of its neighborhood.

Adjacency

If $u \in \Gamma(v)$, u and v are said to be *adjacent*. For undirected graphs, this implies $v \in \Gamma(u)$.

A graph is easily represented by its *adjacency matrix* $\mathbf{A}$: for $V = \{v_1, v_2, \ldots, v_n\}$, element a_{ij} of $\mathbf{A}$ is one if $(v_i, v_j) \in E$ and zero otherwise:

$$a_{ij} = \begin{cases} 1 \text{ if } (v_i, v_j) \in E \\ 0 \text{ if } (v_i, v_j) \notin E \end{cases} \tag{1}$$

Degree

The *degree* of a vertex is the number of edges connected to it, (i.e., the size of its neighborhood for simple undirected graphs) $\deg(v) = |\Gamma(v)|$.

For a directed graph, the *in-degree* $\deg_{\text{in}}(v)$ of a vertex v is the number of incoming edges (u, v) and the *out-degree* $\deg_{\text{out}}(v)$ respectively the number of outgoing edges (v, w).

The *average degree* in a graph is denoted by $\bar{k}$ and the *maximum degree* by Δ. Note that $\overline{k} = 2m/n = \delta(n-1)$.

Degrees in the Web

The **in-degree** represents in a sense the *popularity* of a website $v \in V$: how many administrators of other websites u_i have chosen to put a hyperlink $\langle u_i, v \rangle$ on their website leading to v.

The **out-degree** in turn represents the *forward connectivity* of a website: how many hyperlinks has the administrator of $v \in V$ decided to put on his site pointing to other sites w_i by hyperlinks $\langle v, w_i \rangle$.

Degree distribution

A **degree sequence** of a graph is a vector $\boldsymbol{d} \in \mathbb{Z}^n$ such that $\boldsymbol{d} = (\deg(v_1), \deg(v_2), \ldots, \deg(v_n))$. This can be made unique by sorting the elements in decreasing (or increasing) order.

The **degree distribution** of a graph is a function $p(k)$ that assigns each $k \in [0, n)$ the probability that an arbitrary vertex $v \in V$ has exactly k neighbors, $\Pr[\deg(v) = k]$. The corresponding cumulative distribution is $P(k) = \Pr[\deg(v) \leq k]$.

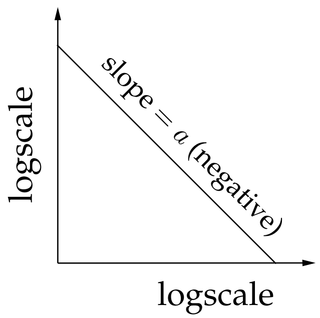
POWER LAWS

A power law is an expression in which a measure y depends on another measure x in linear dependence of the ath power of x, where a is a constant:

$$y \sim x^a.$$

These distributions are also known as *heavy-tail* distributions and include e.g. the Zipfian and Pareto distributions (see e.g. [Sim55, Mit01]).

POWER LAWS ON LOG-LOG PLOTS



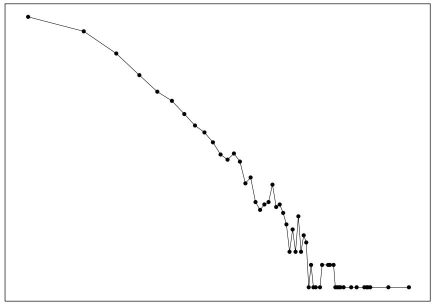
OBSERVED POWER LAWS

In 1999, the Faloutsos brothers [FFF99] reported that several distributions all derived from the topology of the Internet at router (and domain) level followed a power law.

In 2000, also the Web was reported to carry a power law in its link distribution [AH00].

The origins of these power laws have been intensively studied [MMB00, Tad02a, Tad02b].

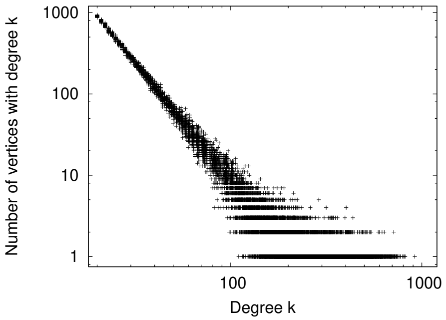
HOW GOOD DOES THE FIT HAVE TO BE?



(Often a good imagination is needed to see the line in research results reported within the complex systems community.)

Scale-free networks

Graphs whose degree distribution follows are power-law (with a negative exponent, denoted by $-a = \gamma$) are generally known as *scale-free* networks (cf. Barabási and Albert [BA99]).



(See [BFS03] pg. 55 for a table of examples.)

Connectivity and paths

Two vertices are *connected* if there exists a *path* $P \subseteq E$ of consecutive edges in the graph between them (in a directed graph, edge orientation is obeyed).

A maximal set of connected vertices in a graph G induces a subgraph that is a *component* of G. A graph is *connected* if all vertices are pairwise connected. If a graph has more than one component, it is *disconnected*.

Path length

Path length is defined as the number of edges on the path, $|P|$. A *simple path* does not visit the same vertex twice.

The *distance* $d\,(u, v)$ between vertices u and v is equal to the length of the shortest path from u to v in G. If no such path exists in G, the distance is infinite (by convention).

Extreme values of path length

The maximum value of $d\,(u, v)$ from a *fixed* vertex u to any other vertex $v \in V$ is called the *eccentricity* of u. The minimum eccentricity over *all* vertices is called the *radius* of the graph G.

The maximum $d\,(u, v)$ for all vertex pairs is called the *diameter* of the graph, $\operatorname{diam}(G)$.

Average path length

The *average path length* $\mathcal{L}(G)$ (a.k.a. the *characteristic path length*) is the average value of the distance $d\,(u,v)$ over all possible pairs $\{u,v\}$:

$$\mathcal{L}(G) = \frac{2}{n(n-1)} \sum_{\{u,v\}\subseteq V} d\,(u,v).$$

For the Web, it was estimated in 1999 that $\mathcal{L} \approx 19$ [AJB99]. To calculate $\mathcal{L}$ for a graph whose degree distribution is known, see [CL02].

Note: in many physics publications, this measure is known as the diameter, contradicting established conventions of graph theory.

Induced subgraphs

A set $S \subseteq V$ induces a **subgraph** $G_S = (S, E_s)$ of $G = (V, E)$ such that

$$E_s = \{(u,v) \in E \mid u, v \in S\}.$$

Also an arbitrary $H = (V', E')$ is a subgraph of G if $V' \subseteq V$ and $E' \subseteq E$ and $(u,v) \in E' \Rightarrow u, v \in V'$, but it is not necessarily induced (i.e., some edges may be “missing”).

Clustering

The clustering coefficient $\mathcal{C}_v \in [0,1]$ of a vertex v is the density of the subgraph induced by $\Gamma(v)$.

The *clustering coefficient* $\mathcal{C}(G)$ of a graph $G = (V, E)$ is the average clustering coefficient of its vertices $v \in V$.

1.1 Power-law size

Huberman and Adamic [HA99] have found that **the number of pages per website follows a power law distribution**.

With a 95-percent confidence interval, they found $\gamma \in [1.647,\ 1.853]$ for one database, and $\gamma \in [1.775,\ 1.990]$ for another.

Studies such as this one are based on large search-engine databases, which are the best sources of Web data although necessarily incomplete.

1.2 Power-law connectivity

Albert, Jeong and Barabási [AJB99] have studied the Web graph (limited to `*.nd.edu`) finding power laws for both the in-degree distribution and the out-degree distribution.

Broder et al. [BKM$^+$00] examined the degree distributions of a much larger sample of the Web graph and verified the presence of power laws for the in-degree and the out-degree of the Web graph, originally reported for smaller experiments such as [BA99], reporting $\gamma_{\text{in}} = 2.09$ and $\gamma_{\text{out}} = 2.72$.

Problems with these estimates

Estimates such as the above power laws are necessarily based on samples or subsets of the Web graph. However, sampling the World Wide Web reliably is nontrivial:

- how to define a "random" web page?
- how to construct a method to obtain one?

Taking a "snapshot" of the Web graph by some sort of a webcrawl is slow if a large portion of the network is mapped. The WWW is continuously changing and therefore the parts of the network covered by early phases of the crawl represent in fact a different version of the graph than the parts covered later. As the time span may be days or even weeks, this cannot be overlooked.

[Rag02]

1.3 Small-world networks

The concept of *small-world networks* was brought up in 1998 by Watts and Strogatz [WS98]. A graph $G = (V, E)$, $|V| = n$ and $|E| = m$, is said to possess the *small-world property* if

- the average path length $\mathcal{L}$ is of similar magnitude than it were in a $G_{n,m}$,
- but the clustering coefficient $\mathcal{C}$ is significantly larger.

The Web was shown to be a small-world network by Adamic [Ada99]; also many other natural networks have been observed to possess the above property to some extent (see e.g. [BT02, MdMLD02]).

PageRank

A is a web page with $\deg_{\text{out}}(A)$ links. It has an incoming link from each of the page B_1, ..., B_n.

The *PageRank* of A is (in the simplest form) [BP98]

$$P_r(A) = (1-d) + d\sum_{i=1}^{n} \frac{P_r(B_i)}{\deg_{\text{out}}(B_i)}$$

where $d \in [0,1]$ is a damping factor (e.g. $d = 0.85$). The PageRanks of all pages form a probability distribution.

1.4 Power law of PageRank

The PageRank is considered to be a “random surfer” rank on how likely is it to hit a page while browsing the web.

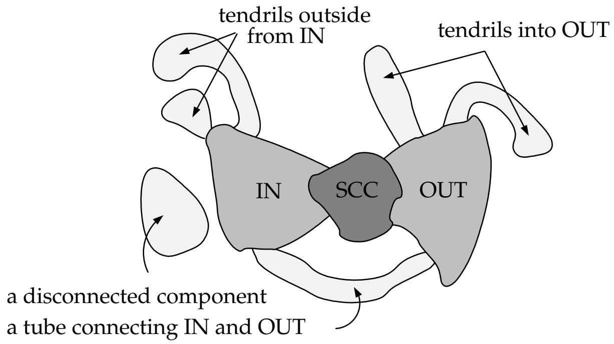
Pandurangan et al. [PRU02] have plotted the distribution of PageRank values for different samples of the Web graph (e.g. the pages under `*.brown.edu` and a 1.69 million document testbed WT10g), finding a power-law distribution with $\gamma = 2.1$.

Structural properties of the Web Graph

A Web sample of 200 million pages and 1.5 billion links was analyzed in 2000 [BKM+00]. The diagram drawn by the authors to represent the structure of the Web Graph has become widely known as the *bow-tie structure*.

A more recent study on how the bow-tie structure evolves is provided in [BYSJC02].

1.5 The bow-tie structure



The SCC contained approx. 56 million pages, IN and OUT both 44 million; the diameter of the SCC at the time was at least 28 (all these are estimates based on random samples).

Hubs and authorities

Kleinberg et al. [KKR+99] observed two special classes of pages in relation to a certain topic in the content of the pages: *authoritative* pages that are devoted contentwise to the topic, and *hub* pages that contain several links to other pages related to the same topic.

A good hub is such that points to good authorities, and a good authority is one that is pointed to by good hubs; the recursive definition is closely related to the PageRank calculation [BP98].

2 Generative Models

There are several stochastic models proposed to mimic the evolution of the Web graph. They differ in their goals as well as foundations; the true evolution of the network must necessarily be simplified, but the simplifications and assumptions made in different models of course vary.

Many of these models were originally proposed to capture the structure of some entirely different network than the Web graph, but have later been viewed as models of the WWW as well, some more justifiably than others.

2.1 Web page growth

According to Baldi [BFS03], Huberman and Adamic proposed the following model for the size $s(t)$ of a single web page at time t:

$$s(t+1) = \alpha(1 + \eta_t \beta)\, s(t),$$

where η is a random coin flip at time t and β represents the *daily fluctuations* of the site size, and α the overall growth rate.

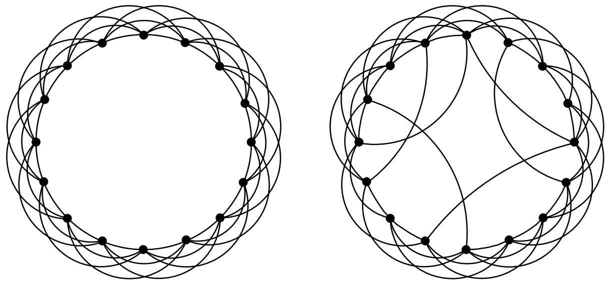
(Baldi cites [HA99], but that is clearly not the correct reference for the above formula.)

2.2 Lattice perturbation models: between order and disorder

Watts and Strogatz [WS98] proposed a simple model to produce networks that exhibit the small-world property: take a ragular lattice, such as the circulant graph (a ring with direct neighborhood connections until a fixed ring-distance), and *rewire* a small fraction of the edges to random nodes instead of the original neighbors.

Also Kleinberg [Kle00b] has proposed such a lattice perturbation model that operates on a square lattice (or a hypercubic lattice in general) and rewires by adding random links with probability that decreases with the lattice (i.e., Manhattan) distance: $\sim d\,(v,w)^{-r}$, where $r > 0$ is constant.

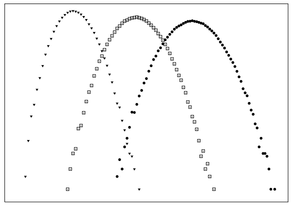
GENERATING SMALL-WORLD NETWORKS



On the left, the circulant graph $C_{16,3}$; a randomly rewired version (with small p) on the right (adapted from [WS98]).

Problems with the rewiring model

First of all, the graph may become disconnected, which can be avoided by a small modification [NW99, NMW00], which also eases the analysis of the model.



The resulting degree distribution is certainly not a scale-free distribution, but many natural networks show both the small-world property and a scale-free degree distribution (see e.g. [ASBS00])

2.3 Preferential attachment models, or the rich get richer

At each time step, add a new vertex v to the graph, and assign d edges to v, selecting the other endpoints among the already existing vertices, with probabilities relative to their degrees:

$$\Pr\left[(v,w) \in E_t\right] = \frac{\deg(w)}{\sum\limits_{u \in V} \deg(u)}.$$

This was originally proposed as the "BA model" of Barabási and Albert [BA99], but the first rigorous mathematical model was defined and analyzed a little later by Bollobás and Riordan [BR00, BRST01].

Problems with the pref. att. models

Simple models that only aim to fulfill the preferential attachment rule fail to capture the small-world phenomenon due to an inherent lack of clustering. Hence hybrid models have been suggested to incorporate more aspects of the observed network topology simulatenously.

Examples omitted here include [DEB02, HK02, LFLW02].

More power-law generators

- variations of preferential attachment [DMS00, ASBS00, VB02]
- by random matchings of "edge stubs" [ACL01]
- deterministic models [BRV01, DGM02, JKK02, RB03]
- generators that match any degree distribution [MR95, MGSZ02]

2.4 Copy models

A graph is generated stepwise by adding new vertices and using random *copying* of neighborhoods from existing vertices to newly added vertices: from an existing vertex $v \in V$, some neighbors $S \subseteq \Gamma(v)$ are chosen to form (a part of) the neighborhood of a new vertex $u \in V$. The above idea of [KKR+99] has been developed further in [KRR+00].

THE RATIONALE BEHIND THE COPY MODEL

Consider the publication of a new web page on some topic with some links to other pages. Some of these links are likely to be the same that are listed on other pages on that topic, such as major organizations related to the topic. This part of the neighborhood structure can be considered as "copied" from some previously existing page.

In addition, the author of the new page will possibly like to contribute another point of view to the topic and is likely to include some "fresh" links on the page in addition to the "established" links on the topic.

2.5 PageRank models

Pandurangan et al. [PRU02], finding that the PageRanks obey a power law, suggest a model in which each new web page links to completely random pages a fraction $\beta \in [0, 1]$ of the time, and otherwise links preferentially to existing pages according to their PageRank values.

They call this model the "PageRank-based selection". However it does not match the observed exponents for the in-degree distribution and the PageRank distribution, although both are power laws. Of course the basic idea allows itself to be modified in several ways to achieve some desired values for exponents of certain distributions.

3 Applications

The observations made regarding the structural properties of networks can be employed in several applications. As examples, we will briefly discuss signaling in complex systems, their robustness to malfunctions, and revealing the community structures underlying the network.

Six degrees of separation

The phrase comes from Stanley Milgram's controversial letter-sending experiment [Mil67, Kle02]: people seem to be able to efficiently construct short chains of acquaintances to connect two random people.

E.g. the Erdős-numbers (`http://www.oakland.edu/~grossman/erdoshp.html`) and the Kevin Bacon graph (`http://www.cs.virginia.edu/oracle/`).

3.1 Distributed search algorithms

Search in power-law and/or small-world networks has been studied by several authors [Wal99, ALPH01, Wal01, KYHJ02, MdM02]. Kleinberg [Kle00a, Kle00b] has shown that the efficient-search regime can be mathematically characterized.

Lada Adamic and collaborators have recently conducted experiments on how messages propagate using different heuristics in social networks; for related work, see [WHAT03].

The performance of peer-to-peer networks can also be improved on the basis of network topology (e.g. by guiding the construction to yield the desired structural properties) [ZGG02].

3.2 Subgraph patterns and communities

In the study of social networks, it is widely believed that there is a community structure in any society: people form communities that are dense in comparison to the connection density between different communities. This topic is brought forward in recent work by Newman and Girvan, but traces back to the 1970s (see [GN02, NG02] and the references therein).

There is a growing interest to discover communities on the Web by analyzing the link structure of the Web graph [FLG00, KL01, FLGC02], and also to identify topics on the Web by similar clustering analysis [WcF00, Men01, GTL^{+}02].

3.3 Robustness and vulnerability

Breakdowns can be imposed on a connected graph by removing some edges or vertices. Vertex removal of course results in the loss of all incident edges as well and is therefore more severe (see e.g. [AJB00]).

Tolerance to *random* breakdowns is called the *robustness* of the network, whereas tolerance to malicious *attacks* is called the *vulnerability*.

Critical threshold p_c

After a sufficient number of node (or edge) removals has taken place, the graph is no longer connected and the size of the largest component is no longer of order n.

The fraction of vertices (together with the incident edges) that need to be removed for this to happen is the critical threshold p_c; the graph remains connected with high probability only if $p \leq p_c$.

The fractions for attack and random breakdown may, and in many interesting cases, do differ significantly.

Properties of scale-free networks

From a scale-free network, an arbitrarily large number of random breakdowns can take place without causing a sharp decrease in the relative size of the largest connected component of the system. Hence the robustness of the network is very good.

However, were the highest-degree vertices eliminated, the network connectivity suffers quickly. Hence the network has poor vulnerability.

Spreading

Studies of dynamical behavior in networks often involve problems of *spreading*, in which the propagation of some *influence* along the network structure is studied. Such influence might be for example a forest fire, heat conduction in metal, or a viral disease.

As different networks allow for different spreading behavior, structural information can be useful in controlling epidemics. For such studies, see e.g. [Mou99, PSV01, MN00, DB02, NFB02, SJ02].

References

[ACL01] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for power law graphs. *Experimental Mathematics*, 10:53–66, 2001.

[Ada99] Lada Adamic. The small world web. In S. Abiteboul and A.-M. Vercoustre, editors, *Proceedings of ECDL'99*, volume 1696 of *Lecture Notes in Computer Science*, pages 443–452, Berlin, Germany, 1999. Springer-Verlag.

[AH00] Lada A. Adamic and Bernardo A. Huberman. Power-law distribution of the World Wide Web. *Science*, 287:2115, March 2000. Techical comment on [BA99] with a response from Barabási et al.

[AJB99] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the World Wide Web. *Nature*, 401:130–131, September 1999.

[AJB00] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, July 2000.

[ALPH01] Lada A. Adamic, Rajan M. Lukose, Amit R. Puniyani, and Bernardo A. Huberman. Search in power-law networks. *Physical Review E*, 64(4):046135, October 2001.

[ASBS00] Luís A. Nunes Amaral, Antonio Scala, Marc Barthelemy, and H. Eugene Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences, USA*, 97(21):11149–11152, October 2000.

[BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.

[BFS03] Pierre Baldi, Paolo Frasconi, and Padhraic Smyth. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*. John Wiley & Sons, Inc., Chichester, UK, 2003.

[BKM+00] Andrei Broder, S. Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer Networks*, 33(1–6):309–320, June 2000.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[BR00] Béla Bollobás and Oliver M. Riordan. The diameter of a scale-free random graph, August 2000. Submitted for publication in Combinatorica, available on-line at `http://www.dpmms.cam.ac.uk/~omr10/`.

[BRST01] Béla Bollobás, Oliver Riordan, Joel Spencer, and Gábor Tusnády. The degree sequence of a scale-free random graph process. *Random Structures and Algorithms*, 18(3):279–290, 2001.

[BRV01] Albert-László Barabási, Erzsébet Ravasz, and Tamás Vicsek. Deterministic scale-free networks. *Physica A*, 299(3–4):559–564, October 15 2001.

[BT02] Tian Bu and Don Towsley. On distinguishing between internet power law topology generators. In *IEEE Infocom: The Twenty-first Annual Joint Conference of the IEEE Computer and Communications Societies*, Los Alamitos, CA, USA, 2002. IEEE Computer Society Press.

[BYSJC02] Ricardo Baeza-Yates, Felipe Saint-Jean, and Carlos Castillo. Web structure, age and page quality. In *String Processing and Information Retrieval*, volume 1627 of *Lecture Notes in Computer Science*, pages 117–130, Berlin, Germany, 2002. Springer-Verlag.

[CL02] Fan Chung and Linyuan Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences, USA*, 99(25):15879–15882, 2002.

[DB02] Zoltán Dezső and Albert-László Barabási. Halting viruses in scale-free networks. *Physical Review E*, 65(5), May 2002.

[DEB02] Jörn Davidsen, Holger Ebel, and Stefan Bornholdt. Emergence of a small world from local interactions: Modeling acquaintance networks. *Physical Review Letters*, 88(12):128701, March 2002.

[DGM02] Sergei N. Dorogovtsev, A. V. Goltsev, and José Ferreira F. Mendes. Pseudofractal scale-free web. *Physical Review E*, 65(6):066122, June 2002.

[DMS00] Sergei N. Dorogovtsev, José Ferreira F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85(21):4633–4636, November 20 2000.

[ER76a] Pál Erdős and Alfréd Rényi. On random graphs i. In *Selected papers of Alfréd Rényi*, volume 2, pages 308–315. Akadémiai Kiadó, Budapest, Hungary, 1976. First publication in Publ. Math. Debrecen 1959.

[ER76b] Pál Erdős and Alfréd Rényi. On the evolution of random graphs. In *Selected papers of Alfréd Rényi*, volume 2, pages 482–525. Akadémiai Kiadó, Budapest, Hungary, 1976. First publication in MTA Mat. Kut. Int. Közl. 1960.

[FFF99] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. In *Proceedings of the ACM SIGCOMM '99 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 251–262, New York, NY, USA, 1999. ACM Press.

[FLG00] Gary William Flake, Steve Lawrence, and C. Lee Giles. Efficient identification of web communities. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, New York, NY, USA, 2000. ACM Press.

[FLGC02] Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35(3):66–71, 2002.

[Gil59] E. N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144, December 1959.

[GN02] Michelle Girvan and Mark E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, USA*, 99:8271–8276, 2002.

[GTL+02] Eric J. Glover, Kostas Tsioutsiouliklis, Steve Lawrence, David M. Pennock, and Gary William Flake. Using web structure for classifying and describing web pages. In *Proceedings of the 11th International World Wide Web Conference*, pages 562–569, New York, NY, USA, 2002. ACM Press.

[HA99] Bernardo A. Huberman and Lada A. Adamic. Growth dynamics of the world-wide web. *Nature*, 401:131–132, September 9 1999.

[HK02] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, 65(2):026107, February 2002.

[JKK02] S. Jung, S. Kim, and B. Kahng. A geometric fractal growth model for scale-free networks. *Physical Review E*, 65(6):056101, May 2002.

[KKR^{+}99] Jon M. Kleinberg, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins. The Web as a graph: Measurements, models, and methods. In T. Asano, H. Imai, D.T. Lee, S. Nakano, and T. Tokuyama, editors, *Proceedings of the Fifth Annual International Conference on Computing and Combinatorics*, volume 1627 of *Lecture Notes in Computer Science*, Berlin, Germany, 1999. Springer-Verlag.

[KL01] Jon Kleinberg and Steve Lawrence. The structure of the web. *Science*, 294(5548):1849–1850, November 2001.

[Kle00a] Jon M. Kleinberg. Navigation in a small world. *Nature*, 406:845, August 24 2000.

[Kle00b] Jon M. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC: Proceedings of the Thirty Second Annual ACM Symposium on Theory of Computing*, pages 163–170, New York, NY, USA, 2000. ACM Press.

[Kle02] Judith Kleinfeld. Six degrees of separation: Urban myth? *Psychology Today*, March/April 2002.

[KRR^{+}00] S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal. Stochastic models for the Web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pages 57–65, Los Alamitos, CA, USA, 2000. IEEE Computer Society Press.

[KYHJ02] Beom Jun Kim, Chang No Yoon, Seung Kee Han, and Hawoong Jeong. Path finding strategies in scale-free networks. *Physical Review E*, 65(2):027103, 2002.

[LFLW02] Mark Levene, Trevor Fenner, George Loizou, and Richard Wheeldon. A stochastic model for the evolution of the web. *Computer Networks*, 39:277–287, 2002.

[MdM02] Cristian F. Moukarzel and M. Argollo de Menezes. Shortest paths on systems with power-law distributed long-range connections. *Physical Review E*, 65(5):056709, May 2002.

[MdMLD02] Adilson E. Motter, Alessandro P. S. de Moura, Ying-Cheng Lai, and Partha Dasgupta. Topology of the conceptual network of language. *Physical Review E*, 65(6):065102, June 2002.

[Men01] Filippo Menczer. Links tell us about lexical and semantic web content. Technical Report CS.IR/0108004, arXiv.org, August 2001.

[MGSZ02] Milena Mihail, Christos Gkantsidis, Amin Saberi, and Ellen Zegura. On the semantics of Internet topologies. Technical Report GIT-CC-02-07, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA, 2002.

[Mil67] Stanley Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[Mit01] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. In *Proceedings of the 39th Annual Allerton Conference on Communication, Control, and Computing*, pages 182–191, Illinois, USA, 2001. University of Illinois.

[MMB00] Alberto Medina, Ibrahim Matta, and John Byers. On the origin of power laws in Internet topologies. *ACM Computer Communication Review*, 30(2):18–28, April 2000.

[MN00] Cristopher Moore and Mark E. J. Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61(5):5678–5682, May 2000.

[Mou99] Cristian F. Moukarzel. Spreading and shortest paths in systems with sparse long-range connections. *Physical Review E*, 60(6):R6263, December 1999.

[MR95] Mike Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6:161–180, 1995.

[NFB02] Mark E. J. Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, September 2002.

[NG02] Mark E. J. Newman and Michelle Girvan. Mixing patterns and community structure in networks. In *Proceedings of the XVIII Sitges Conference on Statistical Mechanics*, Lecture Notes in Physics, Berlin, Germany, 2002. Springer-Verlag. To appear.

[NMW00] Mark E. J. Newman, Cristopher Moore, and Duncan J. Watts. Mean-field solution of the small-world network model. *Physical Review Letters*, 84(14):3201–3204, April 3 2000.

[NW99] Mark E. J. Newman and Duncan J. Watts. Scaling and percolation in the small-world network model. *Physical Review E*, 60(6):7332–7342, 1999.

[PRU02] Gopal Pandurangan, Prabhakar Raghavan, and Eli Upfal. Using PageRank to Characterize Web Structure. In *Proceedings of the Eighth Annual International Computing and Combinatorics Conference*, volume 2387 of *Lecture Notes in Computer Science*, Berlin, Germany, 2002. Springer-Verlag.

[PSV01] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2001.

[Rag02] Prabhakar Raghavan. Discussion at the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, December 2002. Workshop on Topics in Computer Communication and Networks.

[RB03] Erzsébet Ravasz and Albert-Laszlo Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112, February 14 2003.

[Sim55] Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, December 1955.

[SJ02] Pontus Svenson and Desmond A. Johnston. Damage spreading in small world ising models. *Physical Review E*, 65(3):036105, March 2002.

[Tad02a] Bosiljka Tadić. Growth and structure of the world-wide web: Towards realistic modeling. *Computer Physics Communications*, 147(1–2):586–589, August 1 2002.

[Tad02b] Bosiljka Tadić. Temporal fractal structures: origin of power laws in the world-wide web. *Physica A*, 314(1–4):278–283, November 1 2002.

[VB02] Dmitri Volchenkov and Philippe Blanchard. An algorithm generating random graphs with power law degree distributions. *Physica A*, 315(3–4):677–690, December 1 2002.

[Wal99] Toby Walsh. Search in a small world. In *IJCAI'99: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1172–1177, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers.

[Wal01] Toby Walsh. Search on high degree graphs. In *IJCAI'01: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 266–274, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers.

[WcF00] Wai-Chiu Wong and Ada Wai chee Fu. Incremental document clustering for web page classification. In *IEEE 2000 International Conference on Information Society in the 21st century: emerging technologies and new challenges*, 2000.

[WHAT03] Fang Wu, Bernardo A. Huberman, Lada A. Adamic, and Joshua R. Tyler. Information flow in social groups, May 12 2003. Presented in the Annual CNLS conference on Networks in Santa Fe.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small world' networks. *Nature*, 393:440–442, June 1998.

[ZGG02] Hui Zhang, Ashish Goel, and Ramesh Govindan. Using the small world model to improve freenet performance. In *IEEE Infocom: The Twenty-first Annual Joint Conference of the IEEE Computer and Communications Societies*, Los Alamitos, CA, USA, 2002. IEEE Computer Society Press.