A Parsimonious Statistical Protocol for Generating Power-Law Networks

Shilpa Ghadge[1] Timothy Killingback[2] Bala Sundaram[3] Duc A. Tran[1]
[1]Department of Computer Science
[2]Department of Mathematics
[3]Department of Physics
University of Massachusetts, Boston, MA 02125
Email: {shilpa.ghadge001, timothy.killingback, bala.sundaram, duc.tran}@umb.edu

***Abstract*—We propose a new mechanism for generating networks with a wide variety of degree distributions. The idea is a modification of the well-studied preferential attachment scheme in which the degree of each node is used to determine its evolving connectivity. Modifications to this base protocol to include features other than connectivity have been considered in building the network. However, schemes based on preferential attachment in any form require substantial information on the entire network. We propose instead a protocol based only on a single statistical feature which results from the reasonable assumption that the effect of various attributes, which determine the ability of each node to attract other nodes, is multiplicative. This composite attribute or fitness is lognormally distributed and is used in forming the complex network. We show that, by varying the parameters of the lognormal distribution, we can recover both exponential and power-law degree distributions. The exponents for the power-law case are in the correct range seen in real-world networks such as the World Wide Web and the Internet. Further, as power-law networks with exponents in the same range are a crucial ingredient of efficient search algorithms in peer-to-peer networks, we believe our network construct may serve as a basis for new protocols that will enable peer-to-peer networks to efficiently establish a topology conducive to optimized search procedures.**

***Index Terms*—Power-law networks, growing random networks, lognormal distribution, peer-to-peer networks, search in power-law networks**

I. Introduction

In the last decade there has been much interest in studying complex real-world networks and attempting to find theoretical models that elucidate their structure. Although empirical networks have been studied for some time; the recent surge in activity is often seen as having started with Watts and Strogatz's paper on "small world networks" [1]. More recently, the major focus of research has moved from small-world networks to "scale-free" networks, which are characterized by having power-law degree distributions [2]. Empirical studies have shown that in many large networks, like the one shown in Fig. 1 and including the World-Wide-Web [3], the Internet [4], metabolic networks [5], protein networks [6], co-authorship networks [7], and sexual contact networks [8], the degree distribution exhibits a power-law tail: that is, if $p(k)$ is the fraction of nodes in the network having degree k (i.e. having k connections to other nodes) then (for suitably large k)

$$p(k) = ck^{-\lambda}, \tag{1}$$

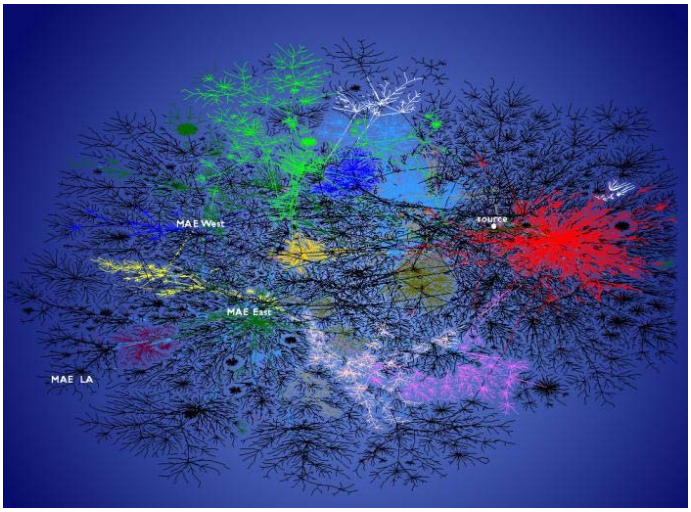

Fig. 1. Diagram in the figure is of Internet connections, showing the major Metropolitan Area Exchanges (MAE), by K.C. Claffy, republished on Albert-László Barabási's Self-Organized Networks Gallery web page "*http://www.nd.edu/ networks/Image Gallery/gallery.htm*". The colors reflect the volume of network traffic.

where $c = (\lambda - 1)m^{\lambda-1}$ is a normalization factor and m is the minimal degree in the network.

One of the earliest theoretical models of a complex network, that of a random graph, was proposed and studied in detail by Erdós and Rényi [9]–[11] in a famous series of papers in the 1950s and 1960s. The Erdós-Rényi random graph model consists of n nodes (or vertices) joined by links (or edges), where each possible edge between two vertices is present independently with probability p and absent with probability $1 - p$.

The degree distribution of the Erdós-Rényi random graph model is easy to determine. The probability $p(k)$ that a vertex in a random graph has exactly degree k is given by the binomial distribution

$$p(k) = \binom{n-1}{k} p^k (1-p)^{n-k-1}. \tag{2}$$

In the limit when $n \gg kz$, where $z = (n-1)p$ is the mean degree, the degree distribution becomes the Poisson

978-1-4244-4581-3/09/$25.00 ©2009 IEEE

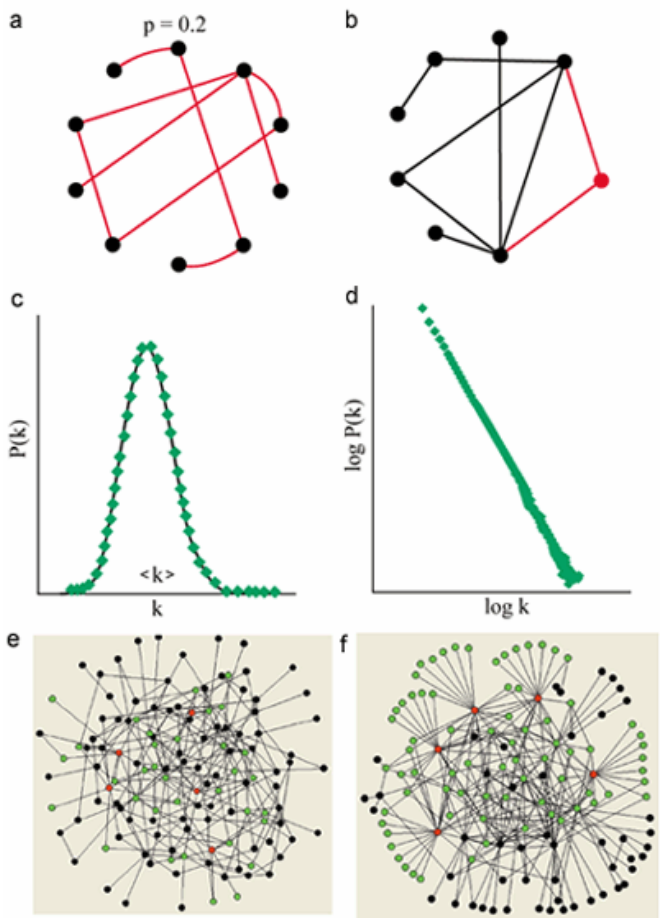

Fig. 2. Construction of random and scale-free networks: the figure is taken from "The physics of the web" by A.L. Barabási, Physics World 14, 33-38 (2001): (a) Erdós Rényi random-graph. The network shown has $N = 10$ and $p = 0.2$.(b) The scale-free construction (c-d) Corresponding degree distributions. (e) The appearance of a random network is rather homogeneous, i.e. most nodes have approximately the same number of links. (f) In scale-free networks a few nodes have a large number of links while the majority are sparsely connected (one or two links). Both the networks shown contain the same number of nodes and links.

distribution

$$p(k) = \frac{z^k e^{-z}}{k!} \tag{3}$$

The Poisson distribution is strongly peaked about the mean z, and has a tail that decays very rapidly as $1/k!$. The rapid decay of the tail of the degree distribution for the Erdós-Rényi random graph is completely different from the heavy-tailed power-law nature of the tail of the degree distribution that is observed in many real-world complex networks. Consequently, random graphs of the Erdós-Rényi type provide poor theoretical models for most real complex networks.

The near ubiquity of heavy-tailed degree distributions (such as the power-law (1)) for real-world complex networks, together with the inadequacy of the Erdós-Rényi random graph as a theoretical model for such networks, brings into sharp relief the fundamental problem of obtaining a satisfactory theoretical explanation for how heavy-tailed degree distribution can naturally arise in complex networks. The dominant concept that is traditionally believed to underlie the emergence of heavy-tailed degree distribution in complex networks is the mechanism of preferential attachment [2]. The preferential attachment algorithm is as follows: starting from a small number (n_0) of nodes (which could for example, be chosen to form a random graph), at every time-step we add a new node with $m \leq n_0$ edges linking the new node to the m distinct nodes already present in the network. The probability Π_i that the new node will connect to node i is taken to be proportional to the degree k_i of that node, thus

$$\Pi_i = \frac{k_i}{\sum_j k_j} \tag{4}$$

This algorithm leads to a growing random network which simulations and analytic arguments show has a power-law degree distribution with $\lambda = 3$. The exponent of the power law is independent of m, which only changes the mean degree of the network. An illustration of the differences in both construction and outcomes of random and scale-free networks is shown in Fig. 2.

Despite the elegance and simplicity of the preferential attachment process it has clear deficiencies as a general explanation for the heavy-tailed nature of the degree distribution of many complex networks. The most basic deficiency of preferential attachment as a mechanism of general importance in real-world networks is that it requires a node that is joining the network to have access to information about the degrees of all the existing nodes. In most real-world situations such information is simply not available, and even in those circumstances in which it is available it is often quite implausible that new nodes use this information when making connections. A brief consideration of one real-world network serves to make the point. The sexual contact network studied in [8] is known to have a power-law degree distribution. The explanation for this phenomenon according to the preferential attachment hypothesis is that a new individual will preferentially seek to have sexual contact with those individuals who already have a large number of sexual contacts. However, the information about how many sexual contacts individuals have is typically not publicly available and cannot be used by any preferential attachment scheme, and even if this information were somehow available the preferential attachment hypothesis seems to be a bizarre explanation for human sexual behavior. Similar objections to the preferential attachment hypothesis are apparent when one considers other complex networks, such as the WWW, or metabolic or protein networks.

II. Lognormal Fitness Attachment Protocol

Here we propose a new explanation for the heavy-tailed nature of the degree distribution of many complex networks that does not require any knowledge of the degrees of existing nodes and is essentially statistical in nature.

To motivate the definition of our procedure we observe that in most real-world networks the nodes will have an attribute

or an associated quantity that represents in some way the likelihood that other nodes in the network will connect to them. For example, in a citation network (see, for example, [7]) the different nodes (i.e. papers) will have different propensities to attract links (i.e. citations). The various factors that contribute to the likelihood of a paper being cited could include the prominence of the author(s), the importance of the journal in which it is published, the apparent scientific merit of the work, the timeliness of the ideas contained in the paper, etc. Moreover, it is plausible that the overall quantity that determines the propensity of a paper to be cited depends essentially multiplicatively on such various factors. The multiplicative nature is likely in this case since if one or two of the factors happen to be very small then the overall likelihood of a paper being cited is often also small, even when other factors are not small. The case of Mendel's work on genetics constitutes an exemplar of this - an unknown author and an obscure journal were enough to bury a fundamentally important scientific paper.

Motivated by this and the other examples we consider it reasonable that in many complex networks each node will have associated to it a quantity, which represents the property of the node to attract links, and this quantity will be formed multiplicatively from a number of factors. That is, to any node i in the network there is associated a non-negative real number Φ_i, which is called the *fitness* of node i, and which is of the form

$$\Phi_i = \prod_{l=1}^{L} \phi_l, \tag{5}$$

where each ϕ_l is non-negative and real. Assuming that the number of factor ϕ_i is reasonably large and that they are statistically independent, then the fitness Φ_i will be lognormally distributed. We recall that a random variable X has a lognormal distribution if the random variable $Y = \ln X$ has a normal distribution. The density function of the normal distribution is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}, \tag{6}$$

where μ is the mean and σ is the standard derivation (i.e. σ^2 is the variance). The range of the normal distribution is $y \in (-\infty, \infty)$. It follows from the logarithmic relation $Y = \ln X$ that the density function of the lognormal distribution is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-(\ln x-\mu)^2/2\sigma^2}. \tag{7}$$

It is conventional to say that the lognormal distribution has parameters μ and σ when the associated normal distribution has mean μ and standard deviation σ. The range of the lognormal distribution is $x \in (0, \infty)$. The lognormal distribution is skewed with mean $e^{\mu+\sigma^2/2}$ and variance $(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$.

The importance of the lognormal distribution comes from the fact that the product of a large number of random variables will be lognormally distributed, irrespective of the manner in which the individual factors are distributed. To be specific, consider a product of M independent random variables $X_1, \ldots, X_M$.

$$X = \prod_{i=1}^{M} X_i. \tag{8}$$

Then

$$\ln X = \sum_{i=1}^{M} \ln X_i \tag{9}$$

and the Central Limit Theorem implies that, essentially irrespectively of how the factors X_i are distributed, the sum will converge to a normal distribution. Thus, $\ln X$ will be normally distributed and X is therefore lognormally distributed. For a general discussion of multiplicative randomness and the lognormal distribution in many areas of science see [12].

Thus, the basic hypothesis that each of the nodes have associated to them a fitness of the form (5), entails that under quite general conditions this fitness will be lognormally distributed. For other approaches to construct networks using fitness associated to the nodes see [13], [14], [15].

We now propose the following procedure to construct a complex network, which we shall refer to as *Lognormal Fitness Attachment*. The protocol is defined as follows: starting from a small number (n_0) of nodes, each of which has associated to it a lognormally distributed random fitness value, at every time-step we add a new node, with its associated lognormally distributed random fitness, and with $m \leq n_0$ edges linking the new node to m distinct nodes already present in the network. The probability Π_i that the new node will connect to node i is taken to be proportional to the fitness Φ_i of that node, thus

$$\Pi(i) = \frac{\Phi_i}{\sum_j \Phi_j}. \tag{10}$$

We note that this scheme does not make any use of information about the degrees of the different nodes. We also note that the only assumption we make about the fitness values is that they are lognormally distributed. We show here that the lognormal fitness attachment model not only results naturally in networks with power-law degree distributions with exponent λ in the range between 2 and 3, which are observed in may real world networks [16], but also in exponential degree distributions, which are also commonly observed, as the single parameter σ is varied.

III. Results

In our construction the fitness parameters Φ_i associated to the nodes in the networks are chosen randomly from a lognormal distribution with parameters μ and σ. As seen from our earlier discussion on lognormal distributions, the standard deviation σ is an important parameter which changes the skewness of the distribution. In order to unambiguously correlate the fitness distribution with properties of the resulting networks, in our calculations, we have kept all other

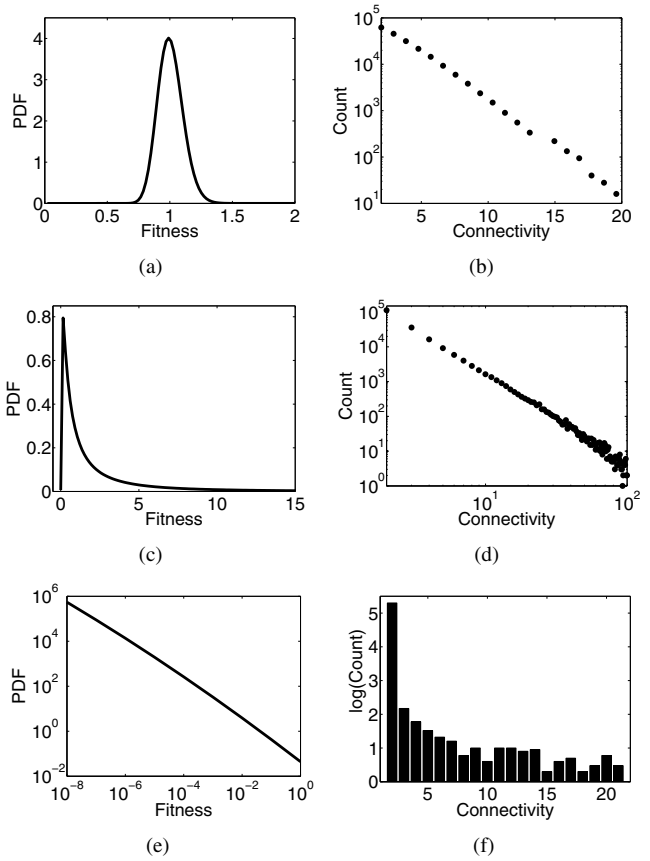

Fig. 3. Changing degree distributions with varying parameter σ for networks with $N = 200,000$ nodes. In each case the side-by-side panels show the fitness distribution and the corresponding degree distribution which results: (a) Lognormally distributed fitness for $\mu = 0$ and $\sigma = 0.1$, (b) Resulting degree distribution corresponding to (a). Note log-linear scale which indicates exponential fall-off. This case would correspond to a random graph; (c) An increasingly skewed distribution with $\mu = 0$ and $\sigma = 1.5$ and (d) corresponding degree distribution plotted on a log-log scale, clearly showing power-law behavior. The corresponding network would be scale-free; (e) Extreme skewed distribution for $\mu = 0, \sigma = 9$. Here the probability of getting a single small value is very high (nearly one), resulting in (f) a distribution where one node completely dominates the network. The fact that the highest degree here is $199,520$ makes clear the common reference to as a 'winner take all' network.

parameters fixed. In each case shown the network is built from a small initial template of node size $n_0 = 3$, each with 2 edges. Each new node comes with two edges, which means $m = 2$. The mean μ is taken to be 0. Though we considered networks of varying sizes N, the results shown correspond to $N = 200,000$. Varying the only remaining parameter σ results in a wide range of network degree distribution properties.

Figure 3 demonstrates the fact that our statistical protocol recovers all recognized network configurations. Three representative values of σ and, in each case, the original fitness distribution and the resulting network degree distribution are shown. In Fig. 3(a) and 3(b), we see the case of small σ ($\sigma = 0.1$). The fitness distribution makes clear the feature that, in this case, the variability in fitness is restricted. This closely resembles a growing random network construction, an expectation borne out by the resulting degree distribution

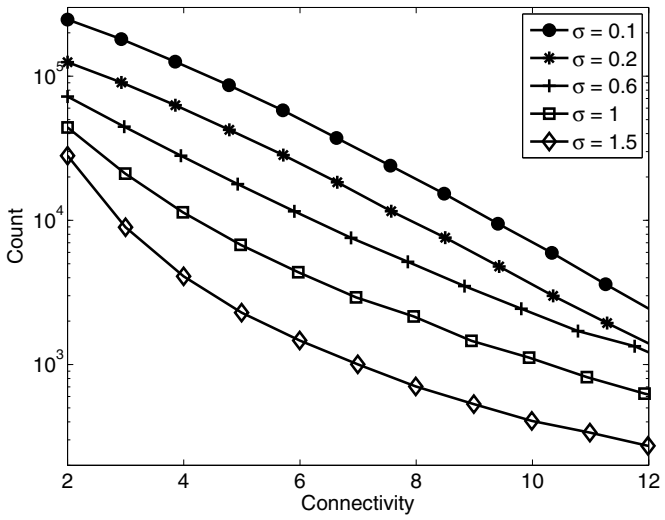

Fig. 4. The seamless evolution from exponential to power-law degree behavior is seen with increasing σ. The fact that the graph is plotted on a log-linear scale makes this transition transparent. The curves have been shifted vertically for clarity.

which exhibits clear exponential behavior.

On making σ larger the fitness distribution allows for a wider range of values which dramatically changes the degree distribution of the resulting network. Here we have a power-law network, which for $\sigma = 1.5$ exhibits an exponent around 3, which is the same as that obtained from the Barabási construction.

On increasing σ to a more extreme value, say 9, we see that most nodes will have small fitness values while very few will get values from the tail. This results in a distribution where a single node dominates the network in terms of connectivity. In the case shown, $199,520$ (out of a possible $199,999$) connections are made to a single node. The next most connected node has a mere 149 connections. This, monopolistic, winner-takes-all outcome has also been noted in some contexts [16].

These results clearly indicate that our protocol recovers the spectrum of networks reported in the literature. Further, the transition between these regimes is not discontinuous, as seen from Fig. 4, where the evolution from random to power-law network behavior is seen with changing σ.

Finally, the power law exponent can also be varied by further changing σ. This is seen from Fig. 5 where, over the range $\sigma = 1.5 - 3.5$, the power law exponent varies from approximately the Barabási value of 3 to 2. Therefore, the scheme we propose has the added advantage of allowing a measure of tunability in the property of the degree distribution of the network. As discussed in the following section, this could prove useful for applications in the context of peer-to-peer networks.

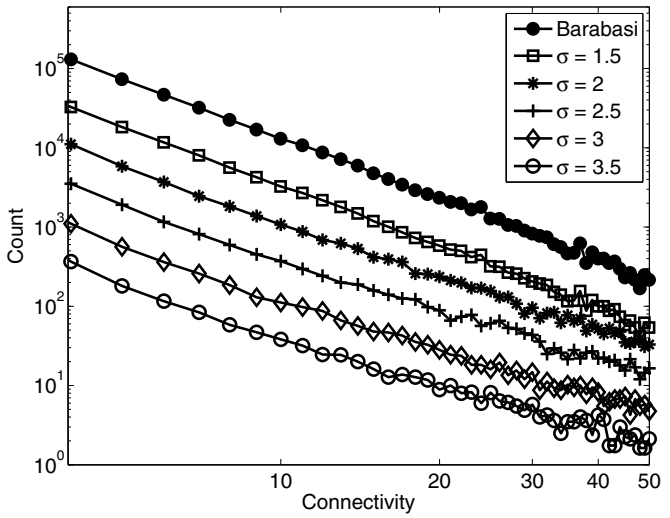

Fig. 5. Changing power-law with increasing σ. These log-log distributions clearly show a changing power-law exponent with a value around 3 for $\sigma = 1.5$ and one closer to 2 for $\sigma = 3.5$. The degree distribution of the Barabási construction, which leads to an exponent of 3, is also shown as a basis for comparison.

σ	2	3	4
Log-normal	17,445	18,696	19,408
Random	18,476	20,393	21,668

TABLE I

MEAN SEARCH TIMES FOR RANDOM-WALK SEARCH ON THREE NETWORKS OF $N = 10,000$ NODES GENERATED USING LOGNORMAL FITNESS ATTACHMENT, WITH $\sigma = 2, 3, 4$, AND THE CORRESPONDING SEARCH TIMES ON THREE REFERENCE RANDOM POWER-LAW NETWORKS WITH THE SAME NUMBER OF NODES AND DEGREE SEQUENCES CONSTRUCTED USING THE CONFIGURATION MODEL (NEWMAN, *et.al*). IN ALL CASES THE MEAN SEARCH TIMES WERE CALCULATED USING $100,000$ RANDOMLY SELECTED START-FINISH PAIRS. THE RESULTS SUPPORT THE CONJECTURE THAT LOGNORMAL FITNESS ATTACHMENT GENERATES A NETWORK WITH A TOPOLOGY MORE AMENABLE TO SEARCH THAN A RANDOM POWER-LAW NETWORK WITH THE SAME DEGREE DISTRIBUTION.

IV. APPLICATION TO PEER-TO-PEER NETWORKS

Peer-to-Peer (P2P) networks consist of a large number of nodes (i.e. computers) that operate in a decentralized fashion to reliably accomplish global tasks for a community of users. Examples of such tasks include database searches and cluster computing. One important class of P2P systems are unstructured P2P networks. Such networks include the file sharing services Gnutella [17], Limewire [18] and Morpheus [19] which together constitute a huge database formed from the dynamical connection of millions of users.

In spite of the popularity of unstructured P2P networks there are major challenges in the systematic design of robust and scalable networks. One of the most important problems is the development of systematic protocols for determining topological characteristic of P2P networks. The importance of this problem comes from the fact that efficient search algorithms for P2P system depend on the network having a power-law structure with exponent between 2 and 3. For a P2P network of size N with such a power-law topology the percolation search algorithm of [20] is capable of finding any content in the network with probability one in time $\mathrm{O}(\log N)$.

It is known from empirical studies [21], [22] that existing P2P networks have approximate power law degree distributions. However, this structure appears to have arisen in an *ad hoc* fashion and, thus, it is important to have client-based protocols that are guaranteed to produce global P2P networks with a predictable power law topology. One such protocol has been proposed in [23]. We believe that our construction of power-law networks with tunable exponents in the range between 2 and 3 provides an alternative basis for a new client-based protocol that guarantees the emergence of a global topology for P2P networks that is suitable for implementing efficient search algorithms, such as that of [20]. The client-level protocol that emerges from our construction of power-law network is straightforward in principle. Each node in a P2P network has a locally generated random variable, the node's fitness, which is drawn from a lognormal distribution, with globally specified values of μ and σ. New nodes entering the network then connect to existing nodes using the lognormal fitness attachment protocol described earlier. As shown, this protocol will automatically result in a P2P network with the global topology of a power law network and with exponent in the range between 2 and 3. As such, this P2P network will therefore be tailored for the percolation search algorithm [20].

In addition to constructing power-law networks with an exponent in the range suitable for efficient search algorithms, we conjecture that the lognormal fitness attachment protocol naturally constructs a power-law network with a global topology that allows more efficient search than a random power-law network with the same degree distribution. This conjecture is supported by the result of random walk search ([24]) on networks generated using our protocol and on random power-law networks with the same degree sequence. Table I shows the results for the mean search time for random walk searches on three networks of $N = 10,000$ nodes constructed using the lognormal fitness attachment protocol and for three reference networks with identical nodes and degree sequences which were constructed using the configuration method [25]. We note that in all cases the random-walk search is more efficient on the networks generated by our protocol than on random networks with the same degree sequence. Testing the efficiency of the percolation search algorithm ([20]) on networks constructed using our protocol is an interesting question currently under consideration.

We note that the lognormal fitness attachment protocol is parsimonious with regard to the information that it requires in the sense that no information concerning the degrees of the different nodes (i.e. the number of computers that are linked to any given computer) is used. We feel that this aspect of the protocol may result in it having particularly efficient implementations.

V. Conclusion

In conclusion, we have suggested a simple statistical method for generating networks with a wide range of degree distributions. A single statistical parameter governs the final outcome which ranges from exponential to power-law to a monopolistic (winner-takes-all) networks. Further, the method does not utilize any information on the connectivity of the existing network structure as the network is built. This protocol naturally constructs power-law networks which are well-suited for efficient search. The protocol may also perhaps be useful in changing the character of an already existing network. These and related issues are currently under consideration.

Acknowledgment

This work was funded in part by the NSF under grants CNS-0615055 and CNS-0753066.

References

[1] D. Watts and S. Strogatz, "Collective dynamics of "small-world" networks," *Nature*, vol. 393, pp. 440–442, 1998.

[2] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[3] R. Albert, H. Jeong, and A. L. Barabási, "Diameter of the world-wide-web," *Nature*, vol. 400, pp. 107–110, 1999.

[4] M. Faloutros, P.Faloutros, and C. Faloutros, "On power-law relationship of the internet topology," *ACM SIGCOM 99, Comp. Comm. Rev.*, vol. 29, pp. 251–260, 1999.

[5] H. Jeong, B. Tombor, R. Albert, N. Oltvai, and A. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 607, p. 651, 2000.

[6] H. Jeong, S. Mason, R. Oltvai, and A. Barabási, "Lethality and centrality in protein networks," *Nature*, vol. 411, p. 41, 2001.

[7] M. Neuman, "The structure of scientific collaboration networks," *Proc. Nat. Acad. Sci. USA*, vol. 98, pp. 404–409, 2001.

[8] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg, "The web of human sexual contacts," *Nature*, vol. 411, p. 907, 2001.

[9] P. Erdós and A. Rényi, "On random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[10] ——, "On the evolution of random graphs," *Publ. of Math. Inst. of the Hungarian Acad. of Sci.*, vol. 5, pp. 17–61, 1960.

[11] ——, "On the strength of connectedness of a random graph," *Acta Mathematia Scientia Hungary*, vol. 12, pp. 261–267, 1961.

[12] E. Limpert, W. A. Stahel, and M. Abbt, "Log-normal distribution across the sciences: keys and clues," *BioScience*, vol. 51-5, pp. 341–352, 2001.

[13] G. Caldarelli, A. Capocci, P. D. L. Rios, and M. Munoz, "Scale-free networks from varying vertex intrinsic fitness," *Phys. Rev. Lett.*, vol. 89-25, p. 258702, 2002.

[14] D. Vito, G. Caldarelli, and P. Butta, "Vertex intrinsic fitness: How to produce arbitrary scale-free networks," *Phys. Rev. E*, vol. 70, p. 056126, 2004.

[15] C. Bedogne and G. J. Rodgers, "Complex growing networks with intrinsic vertex fitness," *Phys. Rev. E*, vol. 74, p. 046115, 2006.

[16] A. Barabási, "The physics of the web," *Physics World*, vol. 14-7, pp. 33–38, 2001.

[17] T. Klingberg and R. Manfredi, "Gnutella protocol development: http://rfc-gnutella.sourceforge.net/src/rfc_06-draft.html," 2002.

[18] C. Rohrs, "Limewire design: http://www.limewire.org/project/www/design.html."

[19] Homepage, "http://www.morpheus.com/index.html."

[20] N. Sarshar, P. Boykin, and V. Roychowdhury, "Percolation search in power law networks: Making unstructured peer-to-peer networks scalable," *Proc. Fourth Int. Conf. on Peer-to-Peer Comp.*, pp. 2–9, 2004.

[21] S. Saroin and S. Gummodi, "A measurement study of peer-to-peer file sharing systems," *Proc. of the Mult. Comp. and Netw.*, 2002.

[22] M. Ripeanu, I. Foster, and A. Iamnitchi, "Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design," *IEEE Inter. Comp. Journal*, vol. 6, p. 2002, 2002.

[23] N. Sarshar and V. Roychowdhury, "Scale-free and stable structures in complex ad hoc networks," *Phys. Rev. E*, vol. 69-2, p. 026101, 2004.

[24] L. Adamic, R. Lukose, A. Puniyani, and B. Huberman, "Search in power-law networks," *Phys. Rev. E*, vol. 64, p. 046135, 2001.

[25] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications," *Phys. Rev. E*, vol. 64, p. 026118, 2001.