The Weibull as a Model of Shortest Path Distributions in Random Networks

Christian Bauckhage
B-IT, University of Bonn
53113 Bonn, Germany

Kristian Kersting
IGG, University of Bonn
53115 Bonn, Germany

Bashir Rastegarpanah
Fraunhofer IAIS
53754, St. Augustin, Germany

ABSTRACT

We address the problem of characterizing shortest path histograms of networks in terms of continuous, analytically tractable distributions. Based on a recent model for the expected number of paths between arbitrary vertices in random networks, we establish the Weibull distribution as the corresponding distribution of minimal path lengths. Empirical tests with different graph topologies confirm our theoretical prediction. Our methodology allows for computing non-linear low dimensional embeddings of path histograms for visual analytics.

Categories and Subject Descriptors

G.2.2 [**Graph Theory**]: Path and Circuit Problems; G.3 [**Probability and Statistics**]: Distribution Functions

General Terms

Theory, Experimentation

Keywords

random networks, shortest path distributions, extreme value theory, Weibull distribution

1. INTRODUCTION

Histograms of shortest path lengths provide useful statistical characterizations of graphs or networks. First of all, features such as average path lengths or graph diameters can be determined therefrom. Second of all, path length statistics are closely related to dynamic properties such as velocities of network spreading processes. Accordingly, if analytical models of shortest path distributions were available, they would facilitate inference and reasoning about network structures and dynamics.

Yet, although the idea of shortest path lengths distributions is an intuitive concept, its analytic treatment proves surprisingly difficult as the combinatorial nature of networks makes often obstructs general results. Related approaches therefore resort to approximations [3, 8].

Here, we extend the work in [3, 8] and draw on *extreme value theory* [5, 16] in order to reason about path length distributions. In particular, we demonstrate that a recent approximation of inter-vertex distances in random networks leads to the Weibull distribution as a model of shortest path lengths statistics. To our knowledge, this characterization has not been provided before. We proceed as follows:

(i) We review a model discussed in [3, 8] and reinterpret it in terms of the expected number of paths between nodes in a random network.

(ii) We summarize key results from extreme value theory and establish the Weibull distribution as an appropriate, physically plausible model of the distribution of shortest path lengths in random networks.

(iii) We present empirical tests that corroborate our theoretical results.

2. DISTANCES IN RANDOM NETWORKS

Following [3, 8], we consider undirected Erdős-Rényi (ER) random graphs $G_{n,\pi}$ of n nodes where any two nodes are connected with probability π. With respect to the distance between a random source node v_i and another randomly selected node v_j, we let F_d denote the probability that the latter is at a distance larger than d from the former. That is, F_d denotes the probability that no path of length less or equal than d exists between v_i and v_j.

Fronczak et al. [8] model F_d for the case of generalized ER graphs. Given two nodes v_i and v_j, they assume modified edge probabilities $\pi_{ij} = h_i h_j / \beta$ where h_i and h_j are node specific hidden variables and $\beta = \mathbb{E}\{h\} n$ is a scaled expectation. They show that F_d can then be written as

$$F_d = e^{-\frac{h_i h_j}{\mathbb{E}\{h^2\} n} \left(\frac{\mathbb{E}\{h^2\} n}{\beta} \right)^d}. \tag{1}$$

Concerned with ordinary ER graphs, Blondel et al. [3] reduce this model to a simpler form. By letting $h_i = np$ for any node v_i in $G_{n,\pi}$, they obtain $\mathbb{E}\{h\} = np$, $\mathbb{E}\{h^2\} = n^2 p^2$, and $\beta = \mathbb{E}\{h\} n$ so that

$$F_d = e^{-\frac{1}{n}(np)^d} \tag{2}$$

where $p = \frac{n-1}{n}\pi$. They interpret this expression in terms of the following recursive process: if a vertex v_j is at a distance larger than d from a randomly chosen source node, all its neighbors are at least $d-1$ steps away from the source. If the number of neighbors is approximated by its expectation np and dependencies are neglected, F_d can be written as

Eleventh Workshop on Mining and Learning with Graphs. Chicago, Illinois, USA
Copyright 2013 ACM 978-1-4503-2322-2 ...$15.00.

$F_d = F_{d-1}^{np}$. The model in (2) is then recovered by setting

$$F_0 = e^{-\frac{1}{n}} \approx 1 - \frac{1}{n}. \tag{3}$$

Next, we provide an alternative interpretation of the model in (2). To this end, we state the following

Theorem 1. *Let $G_{n,\pi}$ be a connected, undirected Erdős-Rényi random graph with n nodes and edge probability π. Let v_i and v_j be any two distinct nodes in $G_{n,\pi}$. The expected number of paths $\mathbb{E}\{N_d\}$ of length $d \geq 2$ between v_i and v_j amounts to*

$$\mathbb{E}\{N_d\} = \frac{1}{n}(np)^d$$

where $p = \frac{n-1}{n}\pi$.

In other words, the exponent in (2) denotes the expected number of paths of length d between any two nodes in an ER graph $G_{n,\pi}$. To prove this, we consider properties of adjacency matrices of undirected ER graphs. The adjacency matrix $\boldsymbol{A}$ of a graph with n nodes is a binary $n \times n$ matrix with entries

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between } v_i \text{ and } v_j \\ 0 & \text{otherwise.} \end{cases}$$

Rows $\boldsymbol{a}^i$ and columns $\boldsymbol{a}_j$ of $\boldsymbol{A}$ therefore are binary vectors. Moreover, as the adjacency matrix of an undirected graph is symmetric, we have $\boldsymbol{A} = \boldsymbol{A}^T$ which implies $\boldsymbol{a}_i^T = \boldsymbol{a}^i$.

Next, recall that $\left(\boldsymbol{A}^d\right)_{ij}$ indicates the number of paths of length d between v_i and v_j. In particular, for $d = 2$, we have

$$\left(\boldsymbol{A}^2\right)_{ij} = \sum_{l=1}^{n} A_{il} A_{lj} = \langle \boldsymbol{a}^i, \boldsymbol{a}_j \rangle = \langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle. \tag{4}$$

That is, the number of paths of length 2 between v_i and v_j is given by the inner product of row $\boldsymbol{a}^i$ and column $\boldsymbol{a}_j$ of $\boldsymbol{A}$ which, as the latter is symmetric, equals the inner product of the corresponding columns.

Finally, recall that if $G_{n,\pi}$ is an ER graph with n nodes and edge probability π, its node degrees are Poisson distributed and the expected node degree is

$$k = (n-1)\pi. \tag{5}$$

An average column $\boldsymbol{a}_i$ of the adjacency matrix $\boldsymbol{A}$ of an ER graph therefore contains k entries equal to 1 which occur with probability $p = \frac{k}{n}$. For column vectors like these, we show

Lemma 1. *Let $\boldsymbol{u}$ and $\boldsymbol{w}$ be two independent n-dimensional binary vectors. If their entries are i.i.d. random variables which equal 1 with probability p and 0 with probability $1-p$, then*

$$\mathbb{E}\{\langle \boldsymbol{u}, \boldsymbol{w} \rangle\} = np^2.$$

is the expected value of the inner product $\langle \boldsymbol{u}, \boldsymbol{w} \rangle$.

Proof. Since the entries u_l and w_l of vectors $\boldsymbol{u}$ and $\boldsymbol{w}$ are independently Bernoulli distributed with

$$P(u_l = b) = P(w_l = b) = p^b(1-p)^{1-b}$$

where $b \in \{0,1\}$, their product $u_l w_l$ is distributed as

$$\begin{aligned} P(u_l w_l = b) &= p^{2b}(1-p^2)^{1-b} \\ &= q^b(1-q)^{1-b} \end{aligned}$$

which is another Bernoulli distribution. The inner product $\langle \boldsymbol{u}, \boldsymbol{w} \rangle = \sum_{l=1}^{n} u_l w_l$ therefore is a sum over n independent Bernoulli trials. Hence, its value is binomially distributed with parameters n and q and its expected value is $nq = np^2$. □

This lemma immediately provides us with an estimate of the expected number of paths of length 2 between any two nodes v_i and v_j in $G_{n,\pi}$, namely

$$\mathbb{E}\{\left(\boldsymbol{A}^2\right)_{ij}\} = \mathbb{E}\{\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle\} = np^2. \tag{6}$$

Using this estimate as a basis for mathematical induction then provides the following

Proof of theorem 1. Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be the adjacency matrix of an ER graph $G_{n,\pi}$. Since the expected degree of any node v is $k = (n-1)\pi$, an average row or column of $\boldsymbol{A}$ contains k entries equal to 1 which occur with probability $p = \frac{k}{n}$. The number of paths of lengths 2 between any two nodes v_i and v_j is given by $\left(\boldsymbol{A}^2\right)_{ij}$ and $\mathbb{E}\{\left(\boldsymbol{A}^2\right)_{ij}\} = np^2$.

Accordingly, assuming independence of $\left(\boldsymbol{A}^2\right)_{il}$ and A_{lj}, the expected number of paths of length 3 between v_i and v_j can be estimated as

$$\begin{aligned} \mathbb{E}\{\left(\boldsymbol{A}^3\right)_{ij}\} &= \mathbb{E}\Big\{\sum_l \left(\boldsymbol{A}^2\right)_{il} A_{lj}\Big\} = \sum_l \mathbb{E}\{\left(\boldsymbol{A}^2\right)_{il} A_{lj}\} \\ &= \sum_l \mathbb{E}\{\left(\boldsymbol{A}^2\right)_{il}\}\,\mathbb{E}\{A_{lj}\} = \sum_l np^2\,\mathbb{E}\{A_{lj}\} \\ &= np^2\,\mathbb{E}\Big\{\sum_l A_{lj}\Big\} \\ &= np^2 np = n^2p^3 \end{aligned}$$

where we have used that $\sum_l A_{lj}$ is the outcome of a series of n Bernoulli trials each with success probability p. Induction leads to $\mathbb{E}\{\left(\boldsymbol{A}^d\right)_{ij}\} = n^{d-1}p^d$ as claimed. □

3. THE WEIBULL MODEL

Next, we show that, for the model of inter-vertex distances in (2), the Weibull distribution naturally arises as a continuous characterization of the distribution of shortest path lengths. First, we summarize key results from extreme value theory and then present theorems that establish our claims.

3.1 Extreme Value Theory

Extreme value theory is concerned with asymptotics of order statistics such as minima or maxima of random samples.

If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables drawn from a distribution with cdf $F(x)$, the cdf of the sample minimum $Y_n = \min_i\{X_i\}$ is found to be

$$F_{Y_n}(y) = P(Y_n \leq y) = 1 - \left(1 - F(y)\right)^n.$$

Since in the limit $n \to \infty$ this distribution is degenerate, extreme value theory studies conditions for non-trivial limiting distributions. The Fisher-Tippett-Gnedenko theorem [7, 10] establishes that there are in fact only three different types of extreme value distributions: (i) the Gumbel, (ii) the Fréchet, and (iii) the Weibull distribution.

The Gumbel distribution arises when $F(x)$ is unbounded from below and has a tail that decreases at least exponentially; the Fréchet distribution arises for distributions $F(x)$ that are unbounded from below and have a tail that declines according to a power law; finally, the Weibull distribution appears if the sampled distribution has a finite lower limit. The latter obviously applies to path lengths in random networks which are lower-bounded by zero.

The pdf and cdf of the standard, two parameter Weibull minimum distribution for $x \geq 0$ are given by

$$f_{\mathcal{WB}}(x; \lambda, \kappa) = \frac{\kappa}{\lambda} \left(\frac{x}{\lambda}\right)^{\kappa-1} e^{-\left(\frac{x}{\lambda}\right)^{\kappa}} \tag{7}$$

and

$$F_{\mathcal{WB}}(x; \lambda, \kappa) = 1 - e^{-\left(\frac{x}{\lambda}\right)^{\kappa}} \tag{8}$$

respectively, where $\lambda > 0$ and $\kappa > 0$ are scale and shape parameters.

The Weibull has the following *minimum closure* property: If $X_1, X_2, \ldots, X_n$ are independent with $X_i \sim f_{\mathcal{WB}}(x; \lambda, \kappa)$ and $Y_n = \min_i\{X_i\}$, then

$$\begin{aligned} F_{Y_n}(y) F_{Y_n}(y) &= 1 - \big(1 - F_{\mathcal{WB}}(y; \lambda, \kappa)\big)^n \\ &= 1 - e^{\left(-\left(\frac{x}{\lambda}\right)^{\kappa}\right)^n} \\ &= 1 - e^{-\left(\frac{x}{\lambda n^{-1/\kappa}}\right)^{\kappa}} \\ &= F_{\mathcal{WB}}\big(y; \lambda n^{-1/\kappa}, \kappa\big). \end{aligned}$$

In other words, if the X_i characterize minima each of which follows a Weibull distribution, then the minimum of the set $\{X_i\}$ is distributed according to another Weibull.

3.2 The Weibull and Shortest Path Lengths

The model in (2) considers ER graphs $G_{n,\pi}$ and expresses the probability F_d for two randomly chosen nodes v_i and v_j being farther apart than d. Accordingly, the expression

$$F(d) = 1 - F_d = 1 - e^{-\frac{1}{n}(np)^d} \tag{9}$$

denotes the probability that v_i and v_j are connected by a path of length less or equal than d. To show that minima of samples drawn from $F(d)$ will be Weibull distributed, we show that $F(d)$ is in the *domain of attraction* of the Weibull.

Theorem 2. *Let $G_{n,\pi}$ be an ER graph as in theorem 1 and assume the validity of the model in (2). Then, $1 - F_d$ is in the domain of attraction of the Weibull distribution so that minima of samples drawn from $1 - F_d$ are Weibull distributed. That is, minimum distances between any two nodes v_i and v_j are Weibull distributed.*

Proof. Gnedenko [10] has shown that a distribution $F(x)$ belongs to the domain of attraction of the Weibull, if the following two criteria are met:

$$\begin{aligned} C_1 &: x_l = \inf\{x \mid F(x) > 0\} > -\infty \\ C_2 &: \lim_{h \downarrow 0} \frac{F(hx - x_l)}{F(h - x_l)} = x^{\gamma}, \quad \gamma > 0. \end{aligned}$$

For inter-vertex distances in graphs, the lower bound $x_l = 0$ is clearly finite so that C_1 is met. In order to verify that C_2 holds for $F(d)$ as defined in (9), we apply l'Hôpital's rule and consider the limiting process

$$\begin{aligned} \lim_{h \downarrow 0} \frac{F(hd - x_l)}{F(d - x_l)} &= \lim_{h \downarrow 0} \frac{\frac{\partial}{\partial h} F(hd - x_l)}{\frac{\partial}{\partial h} F(d - x_l)} \\ &= \lim_{h \downarrow 0} \frac{x\ (np)^{hd - x_l}\ e^{-\frac{1}{n}(np)^{hd - x_l}}}{(np)^{h - x_l}\ e^{-\frac{1}{n}(np)^{h - x_l}}} \\ &= x\ \frac{(np)^{-x_l}\ e^{-\frac{1}{n}(np)^{-x_l}}}{(np)^{-x_l}\ e^{-\frac{1}{n}(np)^{-x_l}}} \\ &= x^{\gamma} \end{aligned}$$

where $\gamma = 1$. Therefore, C_2 is met as well. □

Given our discussion and results so far, distributions of shortest paths in ER networks can be characterized using the following

Theorem 3. *Let $G_{n,\pi}$ be an ER graph as in theorem 1 and assume the validity of the model in (2), then*
(i) the distribution of shortest path between a particular node v_i and a set of nodes $\{v_j\}_{j \neq i}$ follows a Weibull distribution and
(ii) the distribution of all shortest paths in $G_{n,\pi}$ follows a Weibull distribution.

Proof. Both claims follow from theorem 2 together with the minimum closure property of the Weibull distribution. □

3.3 Remarks

To conclude this section, we point out that, even though extreme value theory may now appear as a general analytic tool for treating shortest path distributions, our derivation hinges on properties of ER graphs. Moreover, it hinges on the approximation in (2) with its implicit assumptions as to average node degrees and edge probabilities. If, for instance, the variance of the node degree distribution of a network was too high or a graph had an extreme, non-random layout, e.g. a barbell structure, results concerning expected distances between nodes are much harder to come by (see the discussion in [3]). Nevertheless, our experiments below, in which we consider large sets of Barabási-Albert, power law, and log-normal graphs suggest that for these "natural" graphs, too, shortest path lengths vary in a way that is well accounted for by the Weibull distribution.

Finally, we emphasize that we apply the Weibull as a continuous characterization of discrete distributions; hop counts or path lengths in random network are discrete and their distributions are naturally represented in terms of histograms. Therefore, using the Weibull to represent empirical shortest path histograms is convenient for reasoning and inference but obviously necessitates statistical model fitting.

4. EMPIRICAL EVALUATION

In order to evaluate the merits of the Weibull as a continuous characterization of discrete shortest path distributions, we determined goodness-of-fit results for different kinds of graphs. For baseline comparison, we also considered two alternative models that have been discussed in the related literature.

4.1 Graph Data

We created different Erdős-Rényi (ER), Barabási-Albert (BA), power law (PL), and log-normal (LN) graphs of $n \in \{1,000, 10,000\}$ nodes.

ER graphs are a staple of graph theory and merit investigation. To create ER graphs, we used edge probability parameters $\pi \in \{0.005, 0.0075, 0.01\}$. BA and PL graphs represent networks that result from preferential attachment processes and are frequently observed in biological, social, and technical contexts. To create BA graphs, we considered attachment parameters $m \in \{1, 2, 3\}$ and the exponents of the vertex degree distributions of the PL graphs were drawn from $\gamma \in \{2.1, 2.3, \ldots, 3.1\}$. LN graphs have log-normal vertex degree distributions and reportedly characterize link structures within sub-communities on the web [15]. To synthesize LN graphs, parameters were chosen from $\mu \in \{1, 1.5, \ldots, 3\}$ and $\sigma \in \{0.25, 0.5, 0.75, 1\}$. For each parametrization of all these models, we created 100 graphs, resulting in a total of $64,000$ graphs.

4.2 Baseline Models

An alternative characterization of shortest path length distributions arises from considering a complete graph G with exponentially distributed edge weights. Although G is fully connected, the edge weights effectively thin the graph if we assume that transitions from node to node are more likely for small edge weights. Studying branching processes in graphs like these, Vazquez [18] derives a model in which shortest path lengths follow a Gamma distribution

$$f_{\mathcal{GA}}(x; \theta, \eta) = \frac{1}{\theta^{\eta}} \frac{1}{\Gamma(\eta)} x^{\eta-1} e^{-\frac{x}{\theta}} \tag{10}$$

where $\Gamma(\cdot)$ is the gamma function and $\theta > 0$ and $\eta > 0$ are scale and shape parameters, respectively.

His theoretical prediction is empirically backed by Kalisky et al. [13] who observe that, for spreading trees in scale free networks, the number of nodes per layer is Gamma distributed. However, we point out that Vazquez shows his model only to be valid for PL graphs where $2 < \gamma < 3$; for $\gamma \geq 3$, a different regime takes over.

For additional baseline comparisons, we consider the Log-normal distribution

$$f_{\mathcal{LN}}(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}. \tag{11}$$

In addition to its familiar interpretation in the context of multiplicative growth [14], the Log-normal can be understood as the first passage time distribution of a diffusion with drift [4] and frequently occurs as the distribution of travel times in networks [2, 11].

4.3 Model Fitting and Goodness-of-Fit

We computed the shortest path histogram $\boldsymbol{h}$ of each of our model graphs using Dijkstra's algorithm and applied multinomial maximum likelihood [6, 12] to fit continuous Weibull ($f_{\mathcal{WB}}$), Gamma ($f_{\mathcal{GA}}$), and Log-normal ($f_{\mathcal{LN}}$) distributions. Since statistical tests such as the χ^2 test underestimate the quality of fits to histograms on non-categorial data [9], we used the Kullback-Leibler (KL) divergence

$$D_{KL}(\boldsymbol{h}|\boldsymbol{f}) = \sum_d h_d \log \frac{h_d}{f_d} \tag{12}$$

between empirical data $\boldsymbol{h}$ and fitted model $\boldsymbol{f}$ sampled at d to test goodness of fit. Since the KL divergence measures the loss of information if $\boldsymbol{h}$ is represented by $\boldsymbol{f}$, it follows that the lower the divergence between data and model, the better the model explains the data.

Table 1: Goodness of fit of models fitted to shortest path distributions for graphs of 1,000 nodes.

graph type	parameters	average D_{KL} value		
		$f_{\mathcal{WB}}$	$f_{\mathcal{GA}}$	$f_{\mathcal{LN}}$
Erdős-Rényi	$\pi = 0.005$	**0.009**	0.098	0.343
	$\pi = 0.075$	**0.008**	0.097	0.375
	$\pi = 0.010$	**0.009**	0.079	0.365
Barabási-Albert	$m = 1$	**0.005**	0.020	0.168
	$m = 2$	**0.005**	0.055	0.301
	$m = 3$	**0.012**	0.059	0.344
power law	$\gamma = 2.1$	0.063	**0.008**	0.272
	$\gamma = 2.3$	0.048	**0.006**	0.217
	$\gamma = 2.5$	0.038	**0.007**	0.199
	$\gamma = 2.7$	0.029	**0.010**	0.197
	$\gamma = 2.9$	0.033	**0.011**	0.182
	$\gamma = 3.1$	**0.018**	0.021	0.206
log-normal	$\mu = 1, \sigma = 0.5$	**0.021**	0.084	0.346
	$\mu = 1, \sigma = 1.0$	**0.020**	0.064	0.327
	$\mu = 2, \sigma = 0.5$	**0.016**	0.077	0.366
	$\mu = 2, \sigma = 1.0$	**0.016**	0.054	0.343
	$\mu = 3, \sigma = 0.5$	**0.018**	0.088	0.415
	$\mu = 3, \sigma = 1.0$	**0.019**	0.062	0.392

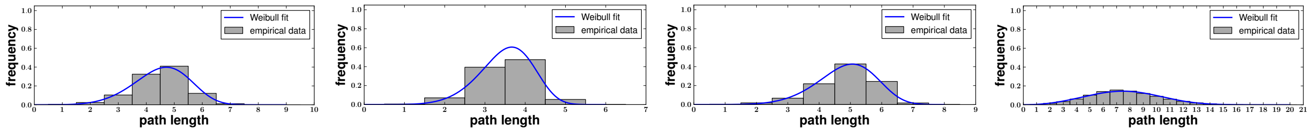
4.4 Results

Table 1 summarizes goodness of fit results for different graphs of $n = 1,000$ nodes in terms of average D_{KL} values. Table 2 lists results for graphs where $n = 10,000$.

Apart from the fact that the average divergence between a fitted model and an empirical distribution is slightly lower for smaller graphs, both tables show strikingly similar results. We observe that (i) in agreement with our theoretical results in section 3, the Weibull distribution provides a well fitting model for the distribution of shortest path lengths in ER graphs; it outperforms the Gamma and the Log-normal distribution; (ii) for BA and for LN graphs, too, the Weibull provides the best fitting model in our tests; (iii) for PL graphs where $2 < \gamma < 3$, the Gamma distribution provides the best model for the distribution of shortest path lengths; this agrees with the theoretical result in [18]; for PL graphs where $\gamma \geq 3$, the Weibull provides the best fitting model; in this context, we note that BA graphs are power law graphs for which $\gamma = 3$ [1]; (iv) for sparser graphs. i.e. for graphs with comparatively fewer edges such as ER graphs where $\pi < 0.01$, BA graphs where $m < 3$, or PL graphs where $\gamma \geq 3$, the Weibull model yields particularly good fits.

Given these empirical results, it appears that the Weibull accounts well for the distribution of shortest path lengths even for topologies different from the ER model.

Figure 1 and 2 illustrate how the Weibull and the Gamma fit empirical shortest path distributions. Visual inspection of many such plots revealed that path length distributions in ER, BA, and LN networks are typically skewed to the left or more or less symmetric. In these cases, the Weibull consistently provided accurate descriptions. For PL networks where $2 < \gamma < 3$, path length distributions were found to be skewed to the right. In these cases, the Gamma achieved more accurate fits. Here, it is interesting to note that both, the Weibull and the Gamma, are special cases of the generalized gamma (GenGamma) distribution [17]. Since the



(a) ER graph, $\pi = 0.005$ (b) BA graph, $m = 3$ (c) LN graph $\mu = 2, \sigma = 0.5$ (d) PL graph $\gamma = 3.1$

Figure 1: Examples of Weibull fits to the shortest path length distributions of an ER, a BA, an LN, and a PL graph. The empirical distributions are either skewed to the left or more or less symmetric; the Weibull mimics this behavior well and closely models the histograms.

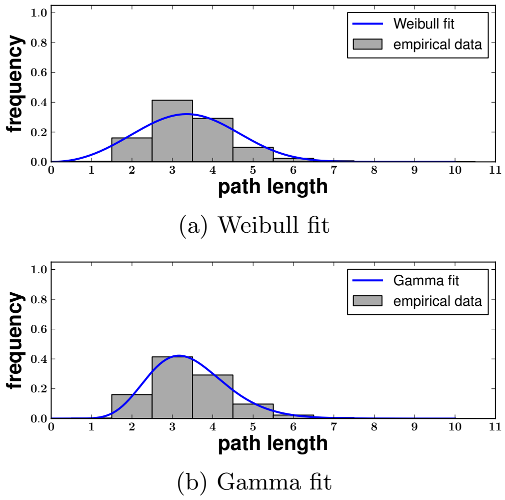
Table 2: Goodness of fit of models fitted to shortest path distributions for graphs of 10,000 nodes.

graph type	parameters	average D_{KL} value		
		$f_{\mathcal{WB}}$	$f_{\mathcal{GA}}$	$f_{\mathcal{LN}}$
Erdős-Rényi	$\pi = 0.005$	**0.014**	0.081	0.240
	$\pi = 0.075$	**0.008**	0.053	0.201
	$\pi = 0.010$	**0.038**	0.057	0.225
Barabási-Albert	$m = 1$	**0.009**	0.017	0.128
	$m = 2$	**0.006**	0.058	0.227
	$m = 3$	**0.015**	0.061	0.248
power law	$\gamma = 2.1$	0.069	**0.008**	0.188
	$\gamma = 2.3$	0.071	**0.005**	0.159
	$\gamma = 2.5$	0.063	**0.007**	0.129
	$\gamma = 2.7$	0.062	**0.015**	0.126
	$\gamma = 2.9$	0.053	**0.027**	0.113
	$\gamma = 3.1$	**0.045**	0.046	0.129
log-normal	$\mu = 1, \sigma = 0.5$	**0.033**	0.099	0.288
	$\mu = 1, \sigma = 1.0$	**0.035**	0.070	0.250
	$\mu = 2, \sigma = 0.5$	**0.023**	0.086	0.281
	$\mu = 2, \sigma = 1.0$	**0.024**	0.054	0.237
	$\mu = 3, \sigma = 0.5$	**0.022**	0.083	0.293
	$\mu = 3, \sigma = 1.0$	**0.027**	0.044	0.240

GenGamma is a three-parameter distribution, it allows for more flexible model fitting than either the Weibull or the Gamma. Accordingly, it seems auspicious to attempt to rigorously unify our theoretical and practical results and those in [18] under the umbrella of the GenGamma. For now, we leave this to future work.

4.5 Embedding Path Length Histograms in 2D

An interesting consequence of using two-parameter distributions to characterize shortest path histograms is that they provide a nonlinear embedding path length data into two dimensions. This allows for visual analytics of the behavior of different graph topologies w.r.t. path length distributions. Figure 3 shows exemplary histograms of shortest path lengths by means of their two-dimensional coordinates (κ, λ) that result from fitting Weibull distributions to the data. Looking at the figure, it appears that shortest path length distributions obtained from different network topologies cluster together or are confined to certain regions in the parameter space. These are preliminary observations which, to the best of our knowledge, have not been reported before. An in-depth study of the characteristics of these representations and possible physical interpretations is underway and results will be reported elsewhere. However, the figure suggests that the idea of characterizing networks in terms of continuous models of shortest path distributions is indeed auspicious and may lead to new insights as to properties of different types of networks..



(a) Weibull fit

(b) Gamma fit

Figure 2: Example of a Weibull and a Gamma fit to the shortest path length distribution of a PL graph ($\gamma = 2.3$). The empirical distribution is noticeably skewed to the right and the Gamma distribution provides the better model.

5. SUMMARY AND FUTURE WORK

We considered the problem of representing discrete path lengths distributions by means of continuous, analytically tractable models. We reinterpreted a recent approximation of inter-vertex distances in random networks in terms of the expected number of paths of length d between two arbitrary nodes. Resorting to extreme value theory, we showed that, for this model, the Weibull distribution naturally arises as the distribution of shortest expected path between nodes.

Empirical tests with different types of random graphs confirmed our theoretical results and revealed that, in addition to Erdős-Rényi graphs, the Weibull also accounts well for shortest path length distributions in Barabási-Albert and Log-normal graphs. For power law graphs, we found the Weibull distribution to provide good fits whenever the power law exponent $\gamma \geq 3$.

As an application in the context of visual analytics, we briefly discussed a non-linear embedding of high-dimensional path length histograms into 2D parameter spaces and observed structural regularities in the resulting representations.

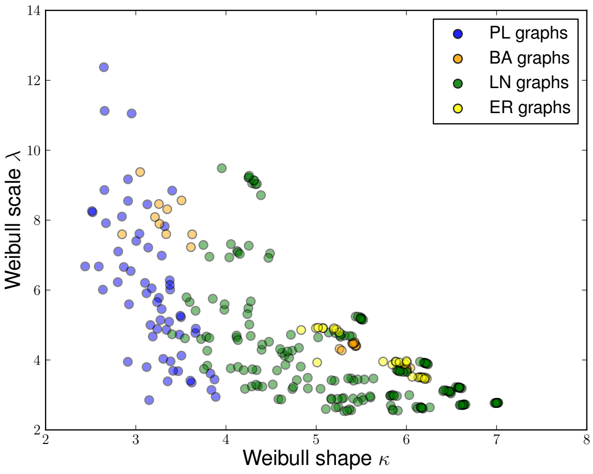


Figure 3: Two-dimensional embedding of shortest path histograms obtained for different kinds of graph topologies. Each point (κ, λ) represents a path length distribution in terms of the parameters of the best fitting Weibull model. Path length distributions obtained from different types of graphs appear to confined to specific regions.

Given these results, there are several directions for future research. First of all, it appears auspicious to unify our results with those of Vazquez in [18] and attempt a characterization of shortest path length histograms in terms of the generalized Gamma distribution which subsumes the Weibull and the Gamma.

Second of all, it appears worthwhile to attempt to connect the shape and scale parameters of the Weibull distribution to physical properties or well established features of networks. Here it seems auspicious to resort to analytical results on expected path lengths in random networks by Fronczak et al. [8]; we expect to be able to establish a connection to, say, the expected value or the variance of the Weibull.

Third of all, we need to extend our approach towards distributions with multiple modes. The obvious strategy is to consider mixtures of Weibull distributions in order to cope with more regular network structures.

Finally, the proposed approach of embedding path lengths histograms in 2D merits further study. If it was possible to establish a connection between locations in these parameter spaces and graph topologies, there will be implications for network inference from outbreak data.

6. ACKNOWLEDGEMENTS

The work reported in this paper was carried out within the Fraunhofer / University of Southampton research project *SoFWIReD* and funded by the Fraunhofer ICON initiative. Kristian Kersting was funded by the Fraunhofer ATTRACT fellowship “Statistical Relational Activity Mining”. The authors gratefully acknowledge this support.

7. REFERENCES

[1] A. Barabasi and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

[2] C. Bauckhage. Insights into Internet Memes. In *Proc. Int. Conf. on Weblogs and Social Media*. AAAI, 2011.

[3] V. Blondel, J. Guillaume, J. Hendrickx, and R. Jungers. Distance Distribution in Random Graphs and Application to Network Exploration. *Physical Review E*, 76(6):066101, 2007.

[4] R. Capocelli and L. Riccardi. On the Inverse of the First Passage Time Probability Problem. *J. Applied Probability*, 9(2):270–287, 1972.

[5] L. de Haan and A. Ferreira. *Extreme Value Theory*. Springer, 2006.

[6] B. Dennis and R. F. Costantino. Ananlysis of Steady State Populations with the Gamma Abundance Model: Application to Tribolium. *Ecology*, 69(4):1200–1213, 1988.

[7] R. Fisher and L. Tippett. Limiting Forms of the Frequency Distribution of the Largest of Smallest Member of a Sample. *Proc. Cambridge Philosophical Society*, 24(2):180–190, 1928.

[8] A. Fronczak, P. Fronczak, and J. Holyst. Average Path Length in Random Networks. *Physical Review E*, 70(5):056110, 2004.

[9] L. Gleser and D. Moore. The Effect of Dependence on Chi-Quare and Empiric Distribution Tests of Fit. *The Annals of Statistics*, 11(4):1100–1108, 1983.

[10] B. Gnedenko. Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire. *Annals of Mathematics*, 44(3):423–453, 1943.

[11] J. Iribarren and E. Moro. Impact of Human Activity Patterns on the Dynamics of Information Diffusion. *Physical Review Letters*, 103(3):038702, 2009.

[12] R. Jennrich and R. Moore. Maximum Likelihood Estimation by Means of Nonlinear Least Squares. In *Proc. of the Statistical Computing Section*. American Statistical Association, 1975.

[13] T. Kalisky, R. Cohen, O. Mokryn, D. Doylv, Y. Shavitt, and S. Havlin. Tomography of Scale-free Networks and Shortest Path Trees. *Physical Review E*, 74(6):077108, 2006.

[14] M. Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[15] D. Pennock, G. Flake, S. Lawrence, E. Glover, and C. Gilles. Winners Don't Take All: Characterizing the Competition for Links on the Web. *PNAS*, 99(8):5207–5211, 2002.

[16] H. Rinne. *The Weibull Distribution*. Chapman & Hall / CRC, 2008.

[17] E. Stacy. A Generalizatin of the Gamma Distribution. *Annals of Mathematics*, 33(3):1187–1192, 1962.

[18] A. Vazquez. Polynomial Growth in Branching Processes with Diverging Reproduction Number. *Physical Review Letters*, 96(3):038702, 2006.