Short Papers

On Modeling Shortest Path Length Distribution in Scale-Free Network Topologies

Agnese V. Ventrella, Giuseppe Piro, *Member, IEEE*, and Luigi Alfredo Grieco, *Senior Member, IEEE*

***Abstract*—Complex and interconnected systems belonging to biological, social, economic, and technology application fields are generally described through scale-free topology models. In this context, it is essential to characterize the distribution of shortest paths in order to obtain precious insights on the network behavior. Unfortunately, the few contributions available in the current scientific literature require a case-by-case tuning of model parameters. To bridge this gap, novel Gaussian-based models are proposed hereby, whose parameters can be immediately tuned based on the number of nodes (N) composing the network, only. In this way, given N, it becomes possible to predict the distribution of shortest paths without retuning the model for each scenario of interest. The outcomes of the proposed models have been successfully validated and compared with respect to state-of-the-art approaches in a wide set of network topologies. To provide a further insight, the conceived Gaussian-based models have been also evaluated for real Internet topologies, learned from reference data sets. Obtained results highlight that the proposed models are able to reach a good tradeoff between the level of accuracy and complexity, even for real network configurations.**

***Index Terms*—Analytical model, internet topology, network theory (graphs), shortest path problem, telecommunication network topology.**

I. Introduction

Scale-free topology models were introduced to catch the logical relationships among entities belonging to a complex and interconnected system. Today, they are widely adopted in biological, social, economic, and technology application fields to describe, for instance, cellular metabolism mechanisms [1], neural networks [2], epidemic phenomena [3], connections among scientific coauthors [4], online social network [5], stocks and shareholders relationships [6], web graphs [7], and network architectures (e.g., autonomous system and overlay nodes [8]). In this context, the shortest path length represents an extremely important parameter, useful to predict the behavior and the performance of the considered interconnected system (think, for example, to the disease spread in a network of people [3], or to the communication latencies in Internet-like topologies [9], and so on). Unfortunately, quite a few contributions investigated the shortest path length between node pairs and the diameter, defined as the maximum distance between any node pair in the topologies. For instance, the distribution of the average shortest path over different topologies has been studied in [7], but the terms "average shortest-path length" and "diameter" have been used as synonyms, which is not the case. This misunderstanding is recurring in the literature: for instance, in [10], the average shortest path length in scale-free networks has been analyzed by referring it as diameter and specifying that it follows a linear distribution (without estimating the coefficient values). Similar considerations apply also to [11]. In [12], several models were investigated, showing that it is possible to catch the distribution of shortest path lengths by properly tuning the parameters of Gamma, Log-normal, and Weibull probability density functions. The main limitation of these approaches is that there is no explicit way to set their parameters. On the contrary, a *case by case* tuning is required: for each scenario of interest, a specific optimal set of parameters has to be used.

This paper intends to complement the current scientific literature by proposing a novel approach that allows us to set the parameters of the shortest path distribution model by *only* considering the number of nodes. In this way, the case-by-case fitting is no longer required. To this end, the network diameter has been first modeled in Section II-A by means of a linear regression. After, starting from the diameter regression, different Gaussian-based models have been proposed and tested to catch the distribution of the shortest path lengths (see Section II-B). A massive simulation campaign has been carried out to clearly demonstrate that the proposed methodology is accurate enough to catch the diameter and the shortest path distribution of scale-free topologies, over a very broad set of conditions and as a function of the number of nodes only. The comparison with respect to the solutions described in [12], further demonstrates that the proposed methodology can be safely adopted because it greatly simplifies the model of the shortest path distribution without any remarkable performance degradation in terms of accuracy. To provide a further insight, the accuracy of the conceived Gaussian-based models that have been also evaluated for real Internet topologies [13] is discussed in Section III. The conducted study further shows that the proposed models register a good tradeoff between the level of accuracy and complexity, even for real network configurations. Finally, Section IV provides closing remarks and draws future research.

II. Analytical Models for Scale-Free Network Topologies

A. Diameter Model

The scale-free model entails an evolving networked system over a discrete time domain: at every timestep, a new vertex is added with $m \leq m_0$ edges, where m_0 is the initial small number of vertices deployed in the system. Typically, little values of m are used to model several systems. For instance, the current literature usually assumes $m = 2$ to model Internet-like topologies [10] and neuronal systems [2]. Also, $m = 1$ and $m = 2$ are used to model the disease spread in a population [3]. Without loss of generality, the rest of this contribution considers $m = 2$. In this case, the *average shortest path*, $\bar{d}$, is approximately

Manuscript received February 14, 2017; revised November 30, 2017 and February 21, 2018; accepted April 1, 2018. This work was supported in part by the research project E-SHELF (by the Apulia Region—Italy, code: OSW3NO1). *(Corresponding author: Giuseppe Piro.)*

A. V. Ventrella is with the Dipartimento di ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Bari 70125, Italy (e-mail: agnese.ventrella@poliba.it).

G. Piro and L. A. Grieco are with the Dipartimento di ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Bari 70125, Italy, and also with Consorzio Nazionale Interuniversitario per le Telecomunicazioni, Parma 43124, Italy (e-mail: giuseppe.piro@poliba.it; alfredo.grieco@poliba.it).

Digital Object Identifier 10.1109/JSYST.2018.2823781

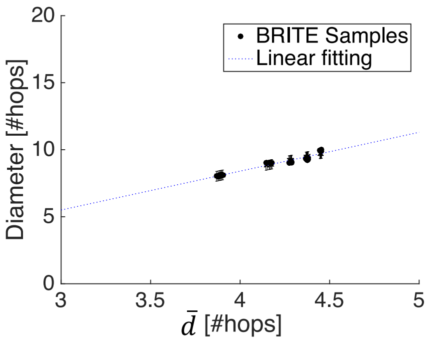


Fig. 1. Diameter linear regression.

equal to [10]:

$$\bar{d} \approx \frac{\log N}{\log \log N} \tag{1}$$

where N is the total number of nodes in the topology.

Starting from (1), the network diameter is modeled herein through a linear regression over $\bar{d}$, as explained below

$$\delta = \alpha \bar{d} + \beta = \alpha \frac{\log N}{\log \log N} + \beta, \tag{2}$$

where α is the y-intercept, β is the slope (or regression coefficient). A massive simulation campaign is carried out to estimate α and β coefficients by means of least-square fitting and their values are obtained as a function of the number of nodes. To this end, BRITE [14] is used to generate different scale-free topologies with a number of nodes (N) ranging from 10^3 to 10^4. For each N, moreover, 50 different topology realizations are evaluated.

The resulting linear regressions are plotted in Fig. 1 (results shows the average diameter over the 50 simulation runs and its confidence interval at 95%).

With reference to results reported in Fig. 1, the least-square fitting procedure has produced the following outcomes: $\alpha = 2.8989$ and $\beta = -3.1938$. Note that the values of α and β coefficients implicitly encompass different network configurations so that the resulting model is able to predict the network diameter as a function of the number of nodes only. In other words, while the linear regression is deducted just one time in this short paper, (2) can be directly used in other contexts, even if the number of nodes and links change, i.e., without requiring any further case by case fitting.

To provide a further insight, the well-known R-squared method is used to evaluate the accuracy of fitting [15]. Accordingly, the R^2 parameter has been evaluated as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\delta_i - \hat{\delta}_i)^2}{\sum_{i=1}^{n} (\delta_i - \bar{\delta})^2},$$

where δ_i, $\hat{\delta}_i$, and $\bar{\delta}$ are the ith average diameter value obtained from simulations for each N, the projection of the aforementioned value to the linear regression calculated according to (2), and the mean of the average diameter samples, respectively. It has been found that $R^2 = 0.92$. Considering that R^2 can be a value in the range $[0, 1]$ and the higher R^2, the better the model, the R-squared method clearly demonstrates that the linear regression is accurate enough to catch the network diameter of scale-free topologies, over a very broad set of number of nodes.

B. Shortest Path Length Model

1) Model Design: Different Gaussian-based models are proposed and tested herein to fit the distribution of shortest path lengths. The probability density function, $f()$, of a generic Gaussian random variable X lying within the interval $X \in (a, b)$, with $-\infty \leq a < b \leq +\infty$, is reported below [15]:

$$f(x; \mu, \sigma, a, b) = \begin{cases} \frac{\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} & \text{if } a \leq x \leq b \\ 0 & \text{elsewhere.} \end{cases} \tag{3}$$

Note that $\phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ and $\Phi(\frac{x-\mu}{\sigma}) = \frac{1}{2}[1 + \text{erf}\,(\frac{x-\mu}{\sigma\sqrt{2}})]$ are the probability density function and the cumulative distribution function (CDF) of the normal distribution with mean μ and standard deviation σ.

Depending on the interval (a, b), four Gaussian-based models can be defined.

Model 1. Gaussian random variable, with $a = -\infty$ and $b = +\infty$. It represents the most general case modeling the shortest paths through a normal distribution.

Model 2. Lower-tail truncated Gaussian random variable, with $a = 0$ and $b = +\infty$. This model imposes that the minimum allowed value of the shortest path is equal to 0.

Model 3. Upper-tail truncated Gaussian random variable, with $a = -\infty$ and $b = \delta$. In this case, the upper bound of the diameter is taken into account to better characterize the shortest path distribution.

Model 4. Two-sided truncated Gaussian random variable, with $a = 0$ and $b = \delta$. This model merges the constraints presented with the previous two cases, thus considering a shortest path distribution lied in the interval $(0, \delta)$.

Considering the (nontruncated) Gaussian-based model, the mean μ is set to $\log N / \log \log N$, in line with the previous section. The standard deviation σ, instead, is estimated by considering the relation between the diameter of the network topology and its average shortest path. Indeed, since the diameter is defined as the longest shortest path length of the topology, it can be represented as the average shortest path $\bar{d}$ plus a certain tolerance interval $k\sigma$, with k being a coefficient of proportionality, as shown below

$$\delta \cong \bar{d} + k\sigma = \frac{\log N}{\log \log N} + k\sigma. \tag{4}$$

Now, by replacing (2) in (4), the standard deviation σ can be rewritten as

$$\sigma = \frac{1}{k} \left(\alpha \frac{\log N}{\log \log N} + \beta - \frac{\log N}{\log \log N} \right). \tag{5}$$

Mean and standard deviation of truncated Gaussian-based models can be finally calculated by using (6) and (7), respectively [15]

$$\bar{\mu} = \mu + \sigma \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} \tag{6}$$

$$\bar{\sigma} = \sigma \sqrt{1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} \right)^2} \tag{7}$$

with $\alpha = \frac{a - \mu}{\sigma}$ and $\beta = \frac{b - \mu}{\sigma}$.

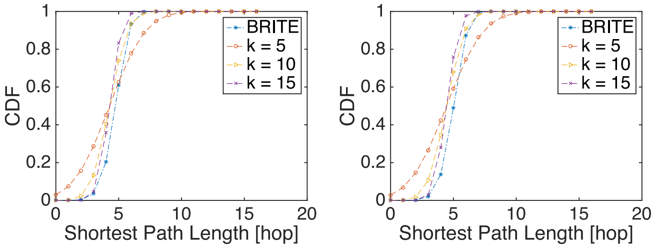
2) Model Tuning: Proposed models are tightly coupled to the coefficient k. For this reason, their CDFs are evaluated hereby for different values of N and k. Fig. 2 reports the CDFs related to Model 1. The curves that refer to the simulated network topologies are shown too.[1] Obtained results clearly demonstrate that the accuracy of proposed models is influenced by k so that it is necessary to undergo a tuning stage. To this end, the theoretical CDFs obtained using the

[1] For lack of space, only results related to networks with 3000 and 30 000 nodes are reported. The missing cases and models exhibit similar behaviors.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
ABSOLUTE ERRORS OF THE 50TH AND 90TH QUANTILES OF THE CDFs OF THE SHORTEST PATHS, EXPRESSED IN TERMS OF NUMBER OF HOPS

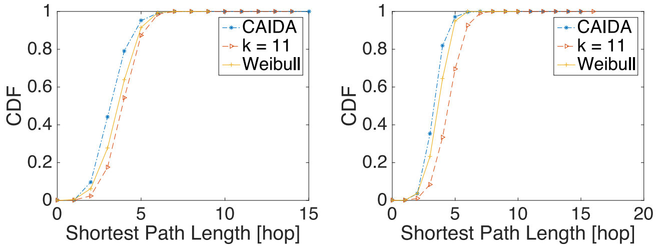
Topologies generated through BRITE												
	Not truncated Gaussian-based						Weibull case by case					
	50th quantiles			90th quantiles			50th quantiles			90th quantiles		
N	3000	10000	30000	3000	10000	30000	3000	10000	30000	3000	10000	30000
Error	0.2721	0.4582	0.5737	0.2539	0.3078	0.4010	0.444	0.389	0.297	0.050	0.015	0.013
Real Internet topologies learned from CAIDA												
	Not truncated Gaussian-based						Weibull case by case					
	50th quantiles			90th quantiles			50th quantiles			90th quantiles		
N	3233	16565	34832	3233	16565	34832	3233	16565	34832	3233	16565	34832
Error	0.7319	0.7199	0.5789	0.4531	0.6401	0.5981	0.4678	0.4429	0.3584	0.2363	0.2180	0.2014



(a)

(b)

Fig. 2. CDF of the shortest path, simulated with BRITE and obtained with Model 1 where (a) N = 3000 and (b) N = 30 000.



(a)

(b)

Fig. 4. CDF of the shortest path, provided by the CAIDA data set and obtained with the proposed Model 1 and Weibull where (a) N = 3233 and (b) N = 34832.

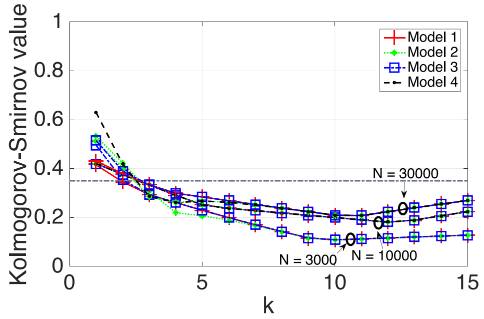


Fig. 3. Outputs of the Kolmogorov–Smirnov test on the overall distributions obtained through BRITE.

four proposed models have been compared with respect to the empirical ones for different values of N and k, using the well-known Kolmogorov–Smirnov test. This test evaluates the maximum absolute difference between the CDFs generated through simulations and the CDFs deriving from the proposed models. Given the number of available samples, a good level of accuracy is reached if the output of the Kolmogorov–Smirnov test is below a threshold equal to 0.35 [16].

From results reported in Fig. 3, it is possible to observe that all the Gaussian-based models provide the same accuracy, given k. In other words, they are equivalent. Moreover, a coefficient $k = 11$ is able to minimize the overall error. According to the Kolmogorov–Smirnov test, in fact, it provides a distance from empirical distributions that lies in the range $[0.112, 0.208]$, below the aforementioned threshold.

3) Performance Comparisons: To assess the degree of accuracy of the proposed models, it is necessary to compare their optimal performance with respect to the models in [12], derived using a case-by-case fitting. To this end, the Kolmogorov–Smirnov test is also executed for the CDF generated with the Weibull model in [12], properly fitted for each considered network topology. In particular, shape and scale parameters of the Weibull distributions, i.e., (γ, λ), are tuned as in what follows: $(4.95, 5.68)$ for $N = 3000$, $(5.54, 6.34)$ for $N = 10\,000$, and $(5.82, 6.52)$ for $N = 30\,000$. The resulting distance between simulated and modeled distributions falls within the range $[0.054, 0.062]$. Therefore, the complexity introduced by the case-by-case fitting provides some performance gain in terms of accuracy with respect to the models proposed in this short paper (i.e., using the case-by-case fitting, it is possible to obtain lower Kolmogorov–Smirnov distances). To quantify the loss of accuracy incurred by the proposed models, the errors associated to the 50th and 90th quantiles of the CDFs of the shortest paths are evaluated and shown in Table I. They clearly demonstrate that the error is always lower than 1 hop. This means that, albeit the proposed models do not require a case-by-case fitting, they are able to predict the median and 90th quantile of the shortest path length with absolute errors less than 1 hop. Indeed, it is possible to conclude that the methodology proposed in this contribution can be safely adopted because it greatly simplifies the model of the shortest path distribution without any remarkable performance degradation in terms of accuracy. In fact, differently from state-of-the-art solutions, it allows to tune the parameters of the shortest path distribution model by *only* considering the number of nodes, without requiring a case-by-case fitting.

III. MODELS VALIDATION IN REAL INTERNET TOPOLOGIES

The accuracy of the conceived Gaussian-based models has been also evaluated for real Internet topologies. Specifically, autonomous system-level topologies are extracted from the CAIDA database [13]. It is important to note that CAIDA does not offer simulated data, but it provides trustworthy topological details of the Internet architecture in different reference years. Thus, considering network topologies from CAIDA is equivalent to conduct real-world topological experiments. The study considered snapshots of the Internet from 1998 to 2010, with a number of node N growing from 3233 to 34 832.

The cumulative distribution of the shortest path of real topologies is reported in Fig. 4, alongside the theoretical CDF related to Model 1. Also in this case, the Kolmogorov–Smirnov test has been executed to evaluate the goodness of the fitting. The outputs are shown in Fig. 5.

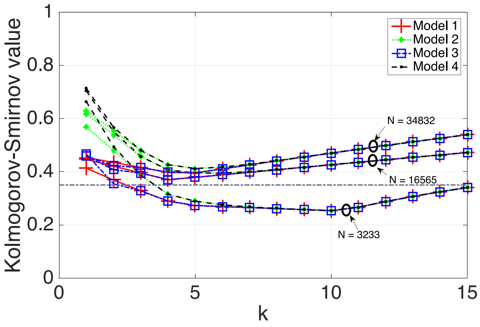


Fig. 5. Outputs of the Kolmogorov–Smirnov test on the overall distributions achieved from the CAIDA data sets.

Given the optimal value of k obtained in the previous section (i.e., $k = 11$), it is possible to observe that the resulting distance between real and modeled distributions falls within the range $[0.246, 0.484]$. On the contrary, the Weibull model registers a distance that falls within the range $[0.165, 0.173]$.

At a first glance, the Kolmogorov–Smirnov test highlights that the Weibull model guarantees the highest level of accuracy and that the proposed models obtain a satisfactory level of accuracy only in real Internet topologies with a limited number of nodes. Anyway, to concretely quantify from a pragmatical perspective, the loss of accuracy incurred by the proposed models, the errors associated to the 50th and 90th quantiles of the CDFs of the shortest paths are reported in Table I. Also, in this case, it emerges that the error is always lower than 1 hop. Therefore, the proposed methodology can be safely adopted also in real Internet topologies, because it offers a good tradeoff between the level of accuracy and the model's complexity.

IV. Conclusion

To the best of the authors' knowledge, this paper demonstrates, for the first time, that it is possible to create simple and effective Gaussian-based models of the shortest path distribution in scale-free network topologies by accounting for the number of nodes only. The accuracy of the proposed models has been further investigated by considering real Internet network topologies, learned from reference data sets. This work represents a strong advancement of the state of the art because currently available approaches require a case-by-case fitting to catch the properties of the network scenario of interest. For future works, we plan to apply this approach to other distributions, such as Weibull, in order to achieve even better accuracy.

References

[1] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[2] K. Fujiwara, T. Tanaka, and K. Nakamura, "Invariant multiparameter sensitivity of oscillator networks," in *Proc. Int. Conf. Neural Inf. Process.*, 2014, pp. 183–190.

[3] M. D. Shirley and S. P. Rushton, "The impacts of network topology on disease spread," *Ecol. Complexity*, vol. 2, no. 3, pp. 287–299, 2005.

[4] Y. Ding, "Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks," *J. Informetrics*, vol. 5, no. 1, pp. 187–203, 2011.

[5] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. SIGCOMM Conf. Internet Meas.*, 2007, pp. 29–42.

[6] D. Garlaschelli, S. Battiston, M. Castri, V. D. Servedio, and G. Caldarelli, "The scale-free topology of market investments," *Physica A: Stat. Mech. Appl.*, vol. 350, no. 2, pp. 491–499, 2005.

[7] R. Albert, H. Jeong, and A.-L. Barabási, "Internet: Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 130–131, 1999.

[8] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *Proc. ACM SIGCOMM Comput. Commun. Rev.*, 1999, vol. 29, pp. 251–262.

[9] J. C. Kuo and W. Liao, "Hop count distribution of multihop paths in wireless networks with arbitrary node density: Modeling and its applications," *IEEE Trans. Veh. Technol.*, vol. 56, no. 4, pp. 2321–2331, Jul. 2007.

[10] B. Bollobás and O. Riordan, "The diameter of a scale-free random graph," *Combinatorica*, vol. 24, no. 1, pp. 5–34, 2004.

[11] R. Cohen and S. Havlin, "Scale-free networks are ultrasmall," *Phys. Rev. Lett.*, vol. 90, no. 5, 2003, Art. no. 058701.

[12] C. Bauckhage, K. Kersting, and B. Rastegarpanah, "The weibull as a model of shortest path distributions in random networks," in *Proc. Workshop Mining Learn. Graphs*, 2013, pp. 1–6.

[13] "Center for Applied Internet Data Analysis (CAIDA)," [Online]. Available: http://www.caida.org/

[14] A. Medina, A. Lakhina, I. Matta, and J. Byers, "BRITE: An approach to universal topology generation," in *Proc. Int. Symp. Model., Anal. Simul. Comput. Telecommun. Syst. Model.*, 2001, pp. 346–353.

[15] N. R. Draper, H. Smith, and E. Pownell, *Applied Regression Analysis*, vol. 3. New York, NY, USA: Wiley, 1966.

[16] E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*. Cambridge, U.K.: Cambridge Univ. Press, 1954.