NIT: A New Internet Topology Generator

Joylan Nunes Maciel[1] and Cristina Duarte Murta[2]

[1] Department of Informatics, UFPR, Brazil
[2] Department of Computing, CEFET-MG, Brazil

Abstract. Internet topology generators play an essential role in computer network research. This paper presents a new Internet autonomous system level topology generator. Synthetic topologies generated by the proposed generator are compared to real Internet topologies and to synthetic topologies generated by a well-known topology generator. The results show that, for all the considered metrics, the proposed generator is able to produce more realistic topologies. This work aims to contribute to the generation of more realistic synthetic topologies.

1 Introduction

The topology of the Internet at the autonomous systems (AS) level has evolved rapidly and its evolution pattern is changing due to network usage, development and deployment. New autonomous systems arise daily and others disappear, and the connections between these systems also change. Recent data [1] reveal that new autonomous systems arise at the rate of 10.3 per day, while the rate of disappearance is 2.87 per day. The connections (links) in this topology arise at a rate of 67.3 per day and disappear at a rate of 45.7 per day. To keep pace with this evolution, the topology must be continuously characterized, the results must be compared with the accepted patterns and reformulated if necessary.

Knowledge acquired in characterizing this evolution is important in many areas of network research including topology analysis. One of the goals of topology mapping is the implementation of topology generators for producing synthetic topologies which are used in network simulators and in laboratory tests [2,3]. The construction of synthetic graphs that represent well the real topology contributes to the effectiveness of tests and experiments of new protocols and Internet applications.

Several Internet topology generators are known, for example Inet [4], GLP [2], BRITE [5], RMAT [6] and Orbis [7]. Others, such as Tiers [8] and Transit-Stub [9], have an important historical character but don't reproduce aspects that are currently observed in the autonomous system topology. Probably the most frequently used autonomous system level topology generator at present is the Inet 3.0, which is also the first recommended on the NS2 network simulator homepage [10]. The current version of Inet (3.0) was built based on the analysis of topologies gathered in the period between November 1997 and February 2002. It has been around since its implementation in 2002.

M. Oliver and S. Sallent (Eds.): EUNICE 2009, LNCS 5733, pp. 108–117, 2009.

This paper introduces NIT, a New Internet Topology generator that is able to generate synthetic topologies that resemble the autonomous systems graph. The need for a new generator is revealed by comparison of actual topologies collected recently with synthetic topologies generated by topology generators. The results indicate discrepancies in many metrics that are essential for topology characterization. The new generator proposed in this paper is capable of generating topologies more similar to real topologies than Inet is, considering the same metrics. The contributions of this paper are the proposal of a new topology generator which is based on Inet code but includes new models for the metrics, and modeling of new metrics; an analysis of the evolution of clique patterns in actual topologies, including maximal clique size and clique size distribution for topologies collected from 2004 to 2007; results of comparison of topologies generated by the new generator proposed in this paper with real topologies recently collected, and synthetic topologies generated by the Inet. The results indicate that the new generator produces topologies that better resemble the real AS-level topologies.

This paper is organized in five sections. Section 2 overviews the well-known topology generators, the topological metrics and the data collection and sources of data of Internet topology. Section 3 presents the new topology generator proposed in this study. Section 4 presents the results of comparisons of synthetic topologies generated by the new generator with real topologies and the synthetic topologies of Inet. The last section concludes the paper.

2 AS-Level Topology: Metrics, Generators, Databases

The Internet does not know its own topology, it has to be discovered. The collection and gathering of data related to the AS level topology have been the subject of a intense debate in the community [3,11,12,13,14,15]. As a consequence, more complete databases have been constructed. Furthermore, the development of the Internet infrastructure and the deployment of new services in the past five years may lead to new evolution patterns.

The topology analysis of complex networks like the Internet can be carried out through a large number of metrics [16,17,18,19]. There is no consensus in the community on the metrics that best represent the topologies of autonomous systems [20,21,14,12]. However, since the discovery that the frequency of node degrees follows a heavy-tailed distribution [22], this metric has been analyzed in several of the cited studies. Besides node degree distribution, the mostly used topological metrics are distances between nodes, distance related metrics, and metrics of graph connectivity such as clique sizes and cluster coefficients.

The discovery of high variability in the distribution of node degrees has motivated the construction of generators based on this characteristic [20,4,5]. To carry out the present study, many topology generators were analyzed with the objective of identifying those that could best reproduce characteristics of the AS-level Internet topology. Initially, we aimed to analyze and compare results from various generators. BRITE, R-MAT, Orbis, GLP and Inet 3.0 generators were considered. But at the end of the analysis, we chose Inet only. The BRITE generator was not included because it had been already compared to Inet, and Inet got better results [4,23]. Moreover, BRITE requires configuration of a large number of parameters, which amplifies the possibilities of topology generation, but makes it extremely difficult to adjust the parameters for a specific topology.

The R-MAT generator [24] recursively subdivides the adjacency matrix that represents the graph into four equal-sized partitions, and connects the partitions using randomly generated edges. Experiments carried out with R-MAT show that it generates non-connected graphs and duplicated edges and self-loops, characteristics which don't occur in the target graph of this study. This fundamental aspect of its design makes it difficult to map parameters of the target topology. Unfortunately, the source code of Orbis and GLP generators were not available at the time of our experiments. The mostly used AS-level topology generator is the Inet 3.0. Its source code is open, its configuration is simple, and it is available on the Web [25].

The AS-level topology database studied in this work was generated by the Internet Research Laboratory of UCLA and is described in [11]. This database, that we call IRL, is the result of a project that seeks to reconstruct the topology of the autonomous systems in a complete and up-to-date manner, collecting additional data from various sources, obtained by diverse methods, including BGP databases, RouteViews and RIPE projects and IRR and Looking Glass databases. The authors have designed and implemented an automatic method for collecting and updating the topology daily. Data is available starting in 2004. We have collected and analyzed eight topologies, one of each semester from 2004 to 2007.

2.1 Overview of Inet 3.0 Topology Generator

Inet relies on the distribution of node degrees as foundation for its structure. The generator's data input is the number of nodes on the graph to be generated. The algorithm is divided into four phases: generation and distribution of the degrees; construction of the connecting component; connection of nodes of degree equal to one; and finally filling edges of nodes having degrees to be completed. The connection of edges is carried out based on the preferential attachment growth model [4].

Inet models the network growth based on the number of nodes, which is related to the "age" of the topology. It begins setting up its baseline in November 1997 when the number of ASes was 3037. This is the minimum number of nodes that Inet accepts as data input. In the first phase, it estimates the number of edges according to the specified number of nodes in the topology, and assigns degrees for all nodes of the graph. The second phase goal is to make one connected component, which is accomplished by generating a spanning tree. In the third phase, the nodes of degree one are connected to the tree, assuring the construction of a single connected component. In the last phase, the lasting connections are done, and at the end all nodes will have a number of connections equal to the degree defined for each one. In Section 4 we present results of analysis of topologies generated by the Inet 3.0.

3 Description of the NIT Topology Generator

In this section we describe the NIT algorithm. The NIT algorithm has five phases: in phase 1 the node degrees are generated; in phase 2 the nodes are connected to build one large connected component of the graph; in phase 3 a number of cliques is generated in the graph; in phase 4, nodes of degree 1 are connected to the graph; and in phase 5 we set the final connections to complete the node degrees. The NIT algorithm is based on the Inet algorithm. NIT conserves the last two phases of the Inet, which are its phases 4 and 5. The first two Inet phases were remodeled in order to better capture the network topology of the present time. Phase 3 of NIT is completely new and it was designed to model the graph connectivity by the insertion of cliques in the topology.

The first task of the algorithm is to size the graph. The number of nodes is the main input but we have to estimate the number of edges based on the degree distribution. We have modeled the degree distribution with the Bounded Pareto distribution which is a heavy-tailed distribution that has minimal and maximal values as parameters. The complement of the Bounded Pareto cumulative distribution function (CCDF) is given by

$$\bar{F}(x) = \frac{j^{\alpha} \times \frac{1}{x^{\alpha}} - (\frac{j}{k})^{\alpha}}{1 - (\frac{j}{k})^{\alpha}} \tag{1}$$

in that parameter j is the minimal value (degree $= 1$) and k is the maximal value, that is obtained by modeling the highest degrees of all topologies considered. Hence, x is an integer variable from $j = 1$ to k. The α parameter models the degree variability and is obtained by the following equation:

$$E[x] = \frac{\alpha}{1 - \alpha} \times \frac{k(\frac{j}{k})^{\alpha} - j}{1 - (\frac{j}{k})^{\alpha}} \tag{2}$$

in that $E[x]$ expresses the average degree, known from the topology data, just as the j and k values. The α values calculated for all real topologies analyzed from 2004 to 2007 as well as our model for this parameter are shown in the plot on the right side of Fig. 1. It is a logarithmic adjustment with coefficient of determination $R^2 = 0{,}999$.

We have also modeled the evolution of the frequency of nodes of degree 1 in all topologies collected between 2004 and 2007. The model is shown in the plot on the left side of Fig. 1. It is modeled by an exponential decay equation that has $R^2 = 0{,}999$.

Therefore, using the models of the percentage of nodes of degree 1 and the parameters of the Bounded Pareto distribution (α, k, and j), we have calculated the distribution of the node degrees and assigned a specific degree to all nodes of the synthetic graph, finishing the first phase of the algorithm.

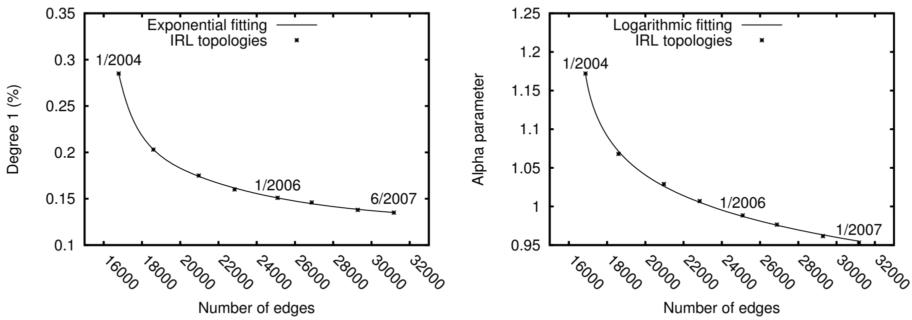

Fig. 1. Fittings for frequency of nodes of degree 1 (left side) and the α parameter (right side) of the IRL topologies

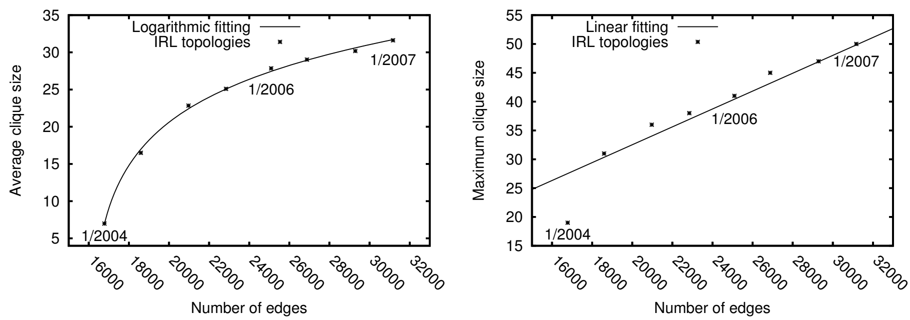

Fig. 2. Fittings for average and maximum clique size of the IRL topologies

The second phase goal is to start the graph as a connected component. This is accomplished by building a spanning tree that includes all nodes of degree three or higher. After that, the nodes of degree 2 are randomly connected. The connections are designed to avoid the large proportion of small cliques that can be found in Inet 3.0, as mentioned in the literature [4]. Small cliques do not contribute to match the clustering coefficients found in the real topology.

In the third phase, our goal is to mimic the cliques pattern founded in the real topology. We have modeled the distribution of clique sizes in the AS-level topology, which follows a Gaussian distribution. The maximum clique size and average clique size were modeled as a function of the number of nodes in the topology, resulting in a logarithmic equation ($R^2 = 0,998$) representing the average clique size and a linear equation ($R^2 = 0,938$) for representing the maximum clique size, as presented in Fig. 2.

Thus, in the third phase we estimate the average clique size and the maximum clique size as a function of the number of nodes of the required topology using these models. Nodes are chosen randomly considering their degrees to make as many cliques as indicated by the models. Modeling the clique distribution is a very important step as the cliques have considerable influence in metrics such as mean distance between nodes, node eccentricity, diameter and clustering coefficients.

Phases 4 and 5 of NIT are similar to the last two phases of Inet. In the next section we present the analysis of metrics comparing the actual topologies given by the IRL database and the synthetic topologies generated by Inet 3.0 and by the NIT algorithm.

4 Comparing Real and Synthetic Topologies

In this section we present results comparing topologies from three sources: the real AS-level topology (IRL) and the synthetic topologies generated by NIT and Inet. The baseline for comparison is the number of nodes, meaning that a comparison is made among topologies with the same number of nodes, which is a measure of the "age" of the network. The topologies are compared according to the following metrics: number of edges, distribution of distances and eccentricities, mean and maximum clique sizes, and cluster coefficients.

The growth of the number of edges in the topologies is shown in Fig. 3. We observe that NIT reproduces better the evolution of the number of edges in the real network, meaning that the Bounded Pareto distribution seems to be a good model. The number of edges generated by Inet is getting severe underestimated for large topologies, whereas NIT overestimate that value by a constant.

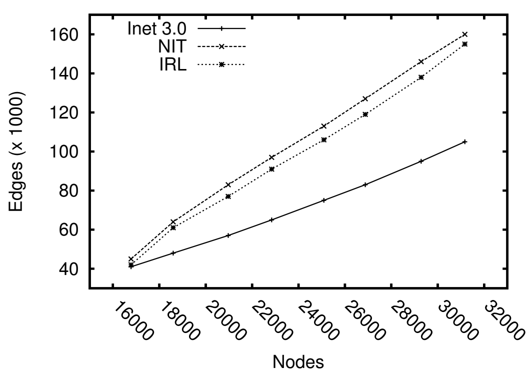

Fig. 3. Edges growth in the Inet 3.0, NIT and IRL topologies

Figure 4 shows the cumulative distribution function (CDF) and the complement of the CDF (CCDF) of vertex degrees for the most recently topology analyzed. We observe that the Bounded Pareto distribution provides a good model for the distribution of node degree. There is a divergence in the tail that includes about 0.1% of nodes but the model of the largest degree helps in the congruence.

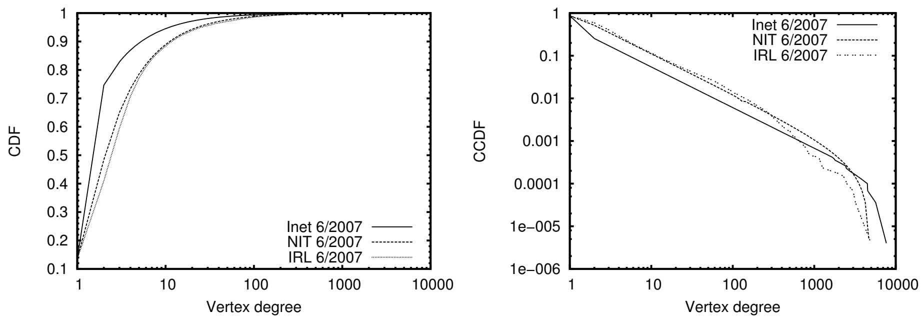

Fig. 4. Degrees Distribution of the Inet 3.0, NIT and IRL topologies

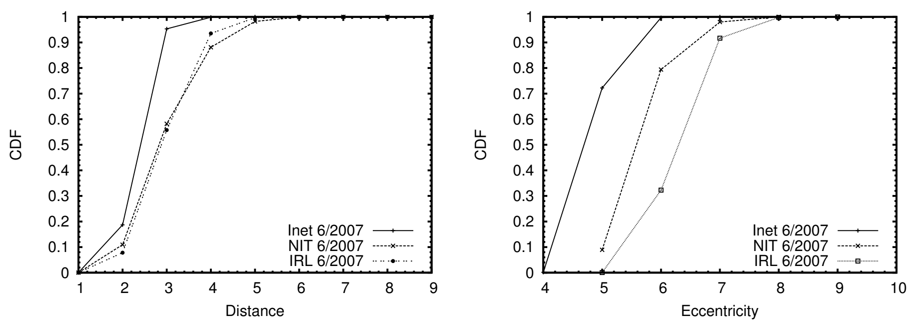

Fig. 5. Distance and Eccentricity Distributions of the Inet, NIT and IRL topologies

The distribution of distances and related metrics are structural metrics of the graph. The distance between two nodes in a graph is the number of edges in the shortest path between them. The eccentricity of a node is the distance to the farthest node. The graph diameter is the maximum eccentricity over all nodes in a graph. The graph radius is the minimum eccentricity over all nodes in a graph.

The cumulative frequency of node distances for all topologies is shown on left plot of Fig. 5. We notice a similar frequency behavior for the real topology and the topology generated by NIT. The good results obtained for the distances are due to the modeling of the number of edges (because more edges helps in graph connectivity), and the design of the clique construction phase (NIT's phase 3), that has influence in the distribution of distances.

The plot on the right of Fig. 5 shows the CDF of the vertex eccentricities in the graph. The analysis shows that the topology generated by NIT presents the same values for diameter and radius of the IRL databases, 9 and 5 respectively, although the frequencies of eccentricities diverge.

Our next analysis focuses on the distribution of clique sizes. Figure 6 shows the CCDF of clique size on the left plot and the evolution of the maximum clique size for all topologies on the right side. We observe that the Inet topology

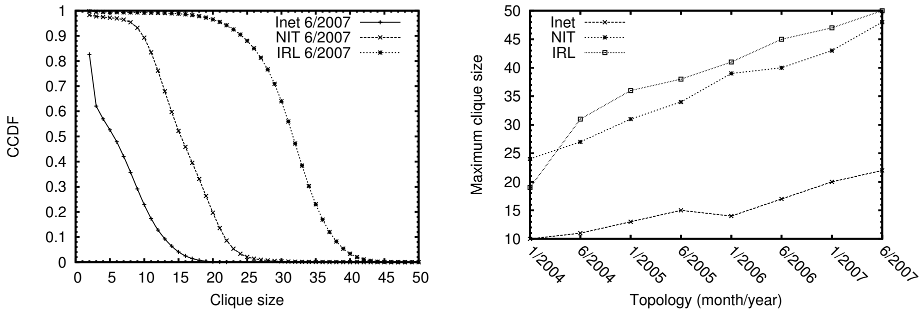

Fig. 6. Clique analysis for all topologies

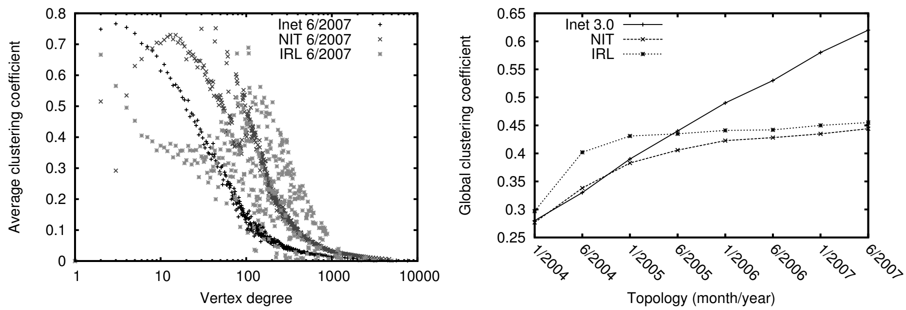

Fig. 7. Clustering coefficients of the Inet, NIT and IRL topologies

generator produces cliques of very small size compared to the real topology. The NIT generator is able to build larger cliques but in a smaller frequency than the real topology, meaning that there is room for improvement.

Our last analysis regards the clustering coefficients. These metrics express the global connectivity of the graph. We have inspected the metrics global clustering coefficient (CC_g) and the average clustering coefficient per degree (CC_m), which are presented on the right and left plots of Fig. 7, respectively. A definition of these metrics can be found in [17].

We observe the variability of CC_m in the real topology compared to the synthetic topologies. The plot on the right shows that the global clustering coefficients of the topologies generated by NIT are getting closer to the values of real topologies.

Finally, we explore the variability of topological metrics generated by different executions of the NIT generator. We generated 100 random topologies of each size (number of nodes) by specifying the first hundred prime numbers as seeds for each execution. We found that the standard deviations for the metric values were very low in all cases. The coefficient of variation, defined as the ratio of the standard deviation to the mean, is lower than 0.06 for all metrics.

5 Conclusion and Future Work

This paper presents a new Internet autonomous system level topology generator called NIT. The need for a new generator is explained by the evolution of the Internet topology in the last years. We have studied and characterized the evolution of the autonomous system level topology in the period from 2004 to 2007. The real topology is growing and getting denser, the distances are shrinking and the average and maximal clique size are becoming larger.

Our generator is based on the Inet 3.0 source code. NIT improves on Inet 3.0 in all metrics tested, introducing new models and a new phase in the algorithm. The results present evidence that the topologies produced by NIT mimics the AS level topology better than the topologies generated by Inet. The usage of topology generators, and particularly Inet, is very significant in computer network research. This work intends to contribute to the generation of more realistic synthetic topologies.

The generation of synthetic topologies that closely resemble the Internet topology is an open problem. The observation that a large number of topologies could be built from a degree distribution [26] turns the approach based on degrees insufficient to model the topology of the Internet. We believe that the association of degree distribution and metrics of distance may establish a new paradigm for the construction of synthetic topology generators, replacing the paradigm of generation based only on the node degree distribution.

Acknowledgements

This research was partially supported by the CNPq and INCT Web. We thank UCLA Internet Research Laboratory for gathering the Internet topology data and making it available on the Web. Joylan thanks LABI/UNIOESTE for supporting his work.

References

1. Oliveira, R.V., Zhang, B., Zhang, L.: Observing the Evolution of Internet AS Topology. SIGCOMM Comput. Commun. Rev., 313–324 (2007)
2. Bu, T., Towsley, D.F.: On Distinguishing between Internet Power Law Topology Generators. In: IEEE INFOCOM, pp. 638–647 (2002)
3. Alderson, D., Li, L., Willinger, W., Doyle, J.C.: Understanding Internet Topology: principles, models, and validation. In: IEEE/ACM Trans. on Networking, pp. 1205–1218 (2005)
4. Winick, J., Jamin, S.: Inet-3.0: Internet Topology Generator. Technical Report 456-02, EECS, University of Michigan (2002)
5. Medina, A., Lakhina, A., Matta, I., Byers, J.W.: BRITE: An Approach to Universal Topology Generation. In: IEEE MASCOTS, pp. 346–361 (2001)
6. Chakrabarti, D., Zhan, Y., Faloutsos, C.: R-MAT: A Recursive Model for Graph Mining. In: SIAM International Conference on Data Mining (2004)

7. Mahadevan, P., Krioukov, D., Fall, K., Vahdat, A.: Systematic Topology Analysis and Generation using degree correlations. In: ACM SIGCOMM, pp. 135–146 (2006)
8. Doar, M.B.: A Better Model for Generating Test Networks. In: IEEE GLOBECOM, pp. 86–93 (1996)
9. Zegura, E.W., Calvert, K.L., Donahoo, M.J.: A quantitative comparison of graph-based models for Internet topology. IEEE/ACM Trans. on Networking, 770–783 (1997)
10. NS2: The Network Simulator NS-2 (2008), `http://www.isi.edu/nsnam/ns/`
11. Zhang, B., Liu, R.A., Massey, D., Zhang, L.: Collecting the Internet AS-level Topology. In: ACM SIGCOMM Computer Communication Review, pp. 53–61 (2005)
12. Mahadevan, P., Krioukov, D., Fomenkov, M., Huffaker, B.: The Internet AS-level Topology: Three Data Sources and One Definitive Metric. In: SIGCOMM Comput. Commun. Rev., pp. 17–26 (2006)
13. Chang, H., Govindan, R., Jamin, S., Shenker, S.J., Willinger, W.: Towards Capturing Representative AS-level Internet Topologies. In: ACM SIGMETRICS, pp. 280–281 (2002)
14. Chang, H., Willinger, W.: Difficulties Measuring the Internet's AS-Level Ecosystem. Invited Paper. In: IEEE INFOCOM, pp. 1479–1483 (2006)
15. Raz, D., Cohen, R.: The Internet Dark Matter - on the Missing Links in the AS Connectivity Map. In: IEEE INFOCOM (2006)
16. Barabási, A.L., Albert, R.: Emergence of Scaling in Random Networks. Science 286, 509–512 (1999)
17. Pastor-Satorras, R., Vespignani, A.: Evolution and Structure of the Internet: A Statistical Physics Approach. Cambridge University Press, Cambridge (2004)
18. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford (2003)
19. Magoni, D., Pansiot, J.J.: Analysis of the Autonomous System Network Topology. In: ACM SIGCOMM Comput. Commun. Rev., pp. 26–37 (2001)
20. Tangmunarunkit, H., Govindan, R., Jamin, S., Shenker, S., Willinger, W.: Network Topology Generators: degree-based vs. structural. In: ACM SIGCOMM, pp. 147–159 (2002)
21. Park, S.T., Pennock, D.M., Giles, C.L.: Comparing Static and Dynamic Measurements and Models of the Internet's AS Topology. In: IEEE INFOCOM, pp. 1616–1627 (2004)
22. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On Power-law relationships of the Internet topology. In: ACM SIGCOMM, pp. 251–262 (1999)
23. Jin, C., Chen, Q., Jamin, S.: Inet: Internet Topology Generator. Technical Report 443-00, Department of EECS, University of Michigan (2000)
24. Chakrabarti, D., Faloutsos, C.: Graph mining: laws, generators, and algorithms. ACM Computing Survey 38(1) (2006)
25. Inet: Inet homepage (2008), `http://topology.eecs.umich.edu/inet/`
26. Li, L., Alderson, D., Willinger, W., Doyle, J.: A First-principles Approach to Understanding the Internet's router-level Topology. In: SIGCOMM Comput. Commun. Rev., pp. 3–14 (2004)