Jellyfish: A Conceptual Model for the AS Internet Topology

Georgos Siganos
U. C. Riverside
siganos@cs.ucr.edu

Sudhir L. Tauro
U. C. Riverside
stauro@cs.ucr.edu

Michalis Faloutsos
U. C. Riverside
michalis@cs.ucr.edu

Abstract

Several novel concepts and tools have revolutionized our understanding of the Internet topology. Most of the existing efforts attempt to develop accurate analytical models. In this paper, our goal is to develop an effective conceptual model: a model that can be easily drawn by hand, while at the same time, it captures significant macroscopic properties. We build the foundation for our model with two thrusts: a) we identify new topological properties, and b) we provide metrics to quantify the topological importance of a node. We propose the jellyfish as a model for the inter-domain Internet topology. We show that our model captures and represents the most significant topological properties. Furthermore, we observe that the jellyfish has lasting value: it describes the topology for more than six years.

1 Introduction

"How can we represent the network graphically in a way that a human can draw or understand?". "How can we define a hierarchy in the Internet topology?"

These are the two main questions that we address in this paper. The overarching goal is to provide a conceptual model for the Internet topology at the Autonomous System (AS) level. Most current research on topology attempts to maintain and describe the information in all its detail. However, a simple conceptual model is also important, especially when it captures graphically many fundamental properties. An example of a successful conceptual model is the bow-tie model used to describe the structure of the world wide web [5].

Conceptual models demonstrate the following paradox: they are difficult to think of, but once they are presented they seem obvious. In our case, the difficulty lies in identifying an "anchor" and a "compass": a well-defined starting point and a way to explore the topology systematically. The main challenge is that the topology is large, complex and constantly changing. Even with the introduction of power-laws, we do not have a comprehensive model of the topology [34][29][21]. Second, although the Internet is widely believed to be hierarchical by construction, it is too interconnected for an obvious hierarchy[35]. Several efforts to visualize the topology have been made [8] [27], but their goal is slightly different from ours: they attempt to show all the available information. In addition, several of those models target the topology at the router-level. These visualizations are useful for multiple different reasons, but they do not meet our requirements: they can not be recreated manually and they do not provide a memorable model.

In this paper, we propose a jellyfish structure as a conceptual model for the Internet topology, extending our work in [36]. The value of the model lies in its simplicity and its ability to capture graphically many topological properties. We use real Internet instances for over six years for our experiments. *First*, we identify a number of new interesting topological properties that

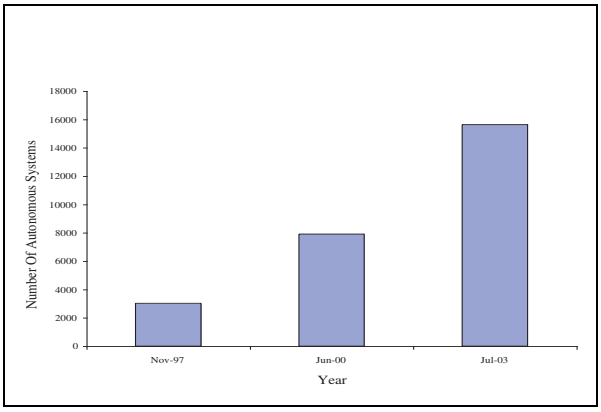

Figure 1: The evolution of the size of the Internet.

guide the development and validate our model. *Second*, we identify metrics for the topological "importance" of a node. We use these metrics to validate our model and establish an anchor: a highly interconnected group of important nodes. *Third*, we show how the topology can be mapped to a jellyfish. *Fourth*, we observe that the jellyfish structure has not changed significantly during more than six years. The network grows "horizontally" by populating its layers, and not by adding new layers. *Finally*, using our model, we show how we can evaluate graph generators. We find that one of the best graph generators fails to capture the macroscopic structure of the Internet.

The rest of this paper is structured as follows. Section 2 presents some background and previous work. Section 3 presents several interesting topological properties, which guide the development and justify our jellyfish model. Section 4 develops a conceptual model for the Internet topology. Section 5 studies the time evolution of the Internet regarding the properties of our model. Section 6 compares the Internet topology with generated topologies. Finally, section 7 concludes our work.

2 Background

The Internet consists of domains or Autonomous Systems (autonomously administered sub-networks of the Internet). The topology of the Internet can be studied at two different levels of granularity. At the router level, we represent each router by a node in the graph. At the inter-domain level, a single node represents each domain and each edge indicates whether the two ASes are directly connected. Here, we study the topology at the inter-domain level or Autonomous System level. We model the topology using an undirected graph.

Definitions and Symbols. The **degree** of a node is defined as the number of edges incident to it. The **distance** between two nodes is the number of edges on a shortest path between the two nodes. The **core** of the graph corresponds to the clique of the highest degree nodes and is defined in more detail in section 4.1. The **effective eccentricity**, $ecc(v)$, of node v is the minimum number of hops required to reach at least 90% of the nodes that are reachable from that node[1]. Note that important nodes have low eccentricity. In the rest of this paper, we refer to effective eccentricity simply as eccentricity. The **significance** of a node attempts to capture both the number and the importance of neighbors. A simple recursive algorithm can be used to calculate the significance [19], which is similar to the page rank notion used by google to rank web pages. Initially, all nodes start with equal significance. At each step, the significance of each node is set to the sum of the significance of its neighbors. At the end of each step, all values are normalized so that their sum equals to one. We stop when the significance of the nodes converges to a set of values. Note that this is equivalent to finding the eigenvector of the maximum eigenvalue of the adjacency matrix of the graph [19]. We define **relative significance** as the product

[1]The effective eccentricity as defined here has already been used successfully to analyze topological properties of the Internet at the router level [28].

of the significance of a node and the number of nodes in the graph[2]. In the rest of this paper, we focus on the relative significance and we use the terms significance and relative significance interchangeably. In order to study the importance of a node we will use the degree, the eccentricity and the significance of a node.

In some cases, we will use power-laws to characterize skewed distributions. A **power-law** is an expression of the form $y \propto x^c$, where c is a constant, x and y are measures of interest and $\propto$ stands for "proportional to". We use linear regression to fit a line to a set of two-dimensional points [30] and the least square errors method. The validity of the approximation is indicated by the correlation coefficient, which is a number between -1 and 1. We refer to the absolute percent value of the correlation coefficient value, for which a value of 100% indicates perfect linear correlation.

Graph Instances. We analyze real instances of the Internet topology from 1997 to 2003. We use the data collected from the Oregon routeviews project [26]. Although the Oregon has been reported to miss edges between ASes [7], it is widely used in many AS studies [27, 13, 14, 35, 7]. The reason for its popularity is that it is the only archival of data that can provide information for the evolution of the topology, and also it captures in a consistent way the nodes of the topology.

The network has grown significantly over the six years of observation. In figure 1, we plot the network size of the three real graphs for most of our experiments. The growth of the Internet in the time period we study is almost 516%. We highlight three instances that we will use more often in this paper:

1. Int-11-97: November of 1997 with 3015 nodes and 5156 edges and 3.42 avg. degree
2. Int-06-2000: June 2000, 7864 nodes and 15713 edges and 3.996 avg. degree.
3. Int-07-2003: July 2003, 15634 nodes and 34689 edges and 4.43 avg. degree.

[2]The definition of relative significance facilitates the interpretation of its value by establishing one as a reference value. If we assume that all nodes have equal importance, then the relative significance values would be equal to one. Therefore, a node with relative significance greater than one has more importance than its "fair share".

Previous work. Modeling the Internet topology has received significant attention recently. However, most of this work does not attempt to develop a conceptual model which is the target of this paper.

Real Network Studies and Properties. Faloutsos et al. [34] identified several power laws that describe concisely distributions of graph properties such as the node degree. Intuitively, their work shows that the topological properties show high variability with few elements having very high values, while the majority of them has below-average values. In an earlier effort, Govindan and Reddy [15] study the growth of the inter-domain topology of the Internet. They classify the domains in a 4-tier hierarchy based on degree. Albert et al. [3] explore the resilience of the network using the average distance between nodes as a metric. Chen et al. [7] express concerns about the completeness and accuracy of the topology from the Oregon project. Tangmunarunkit et al. [35] examine macroscopic topological properties and attempt to develop a framework for comparing topologies.

Theoretical studies. A fascinating study by Reittu and Norros [32] provides theoretical support for our model. Their study proves that graphs with power-law degree distribution and randomly connected nodes (given the power-law degree distribution) will also have the following properties: a) there exists a highly connected core, b) the diameter of the graph is proportional to $\log \log N$. Another study by Cohen and Havlin [9] concurs that small world networks are expected to have small diameter and distances. Our data validates both these theoretically predicted properties. Mihail et al. [24] prove a surprising relationship between the eigenvalues of the adjacency matrix and the degree of the nodes. Recently, there exist a number of theoretical papers [25] [31], [10] that propose the

use of hierarchical network models to characterize graphs with a power-law degree distribution.

Most recently, several theoretical studies on complex networks address the problems of core identification and hierarchy in social and life science networks [37, 39].

Visualization Efforts. There have been few visualization efforts compared to the measurement activity [8, 27, 17]. Most of these efforts attempt to show the entire graph in all its detail. Furthermore, some of these efforts examine the topology at the router level [8, 17].

Graph Generators. We can distinguish Internet models in two categories depending on whether they consider power-laws in their degree distribution. The early graph models assume a uniform degree distribution [38][11]. Zegura et al. [40] introduce a comprehensive model that includes several previous models and combines simple topologies in a hierarchical structure. After the discovery of power-laws, several models have been proposed to capture the skewed degree distribution [22][4][1][6][12][2][18].

Recently, two research efforts study the structure of the logical AS graph, which is a directed graph that represents the business relationships (i.e. customer - provider) apart from the connectivity. Gao et al [14] develop a structural model of the directed AS graph. Subramanian et al. [20] propose a five level classification of ASes based on the commercial relationships.

3 Topological Properties

In this section, we identify several topological properties that provide guidelines for the development of our model. First, we study metrics to quantify the topological importance of a node. Second, we study the spatial distribution of the one-degree nodes in the graph. Third, we study the connectivity of the graph.

In order to study the topological importance [3], we propose three metrics. The degree of a node is a straightforward metric of the importance. Naturally, a high degree suggests higher importance. Additionally, we explore the meaning and the relationships between the eccentricity and the significance.

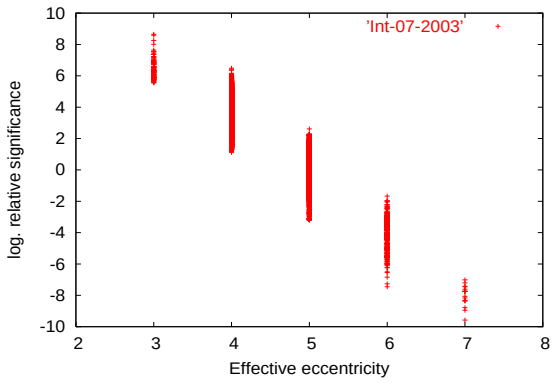

Figure 2: The logarithm of relative significance versus the effective eccentricity for Int-07-2003.

The degree and the significance capture different topological properties. The degree and the significance are related, but at the same time, they capture significantly different aspects of the topology. The degree of a node captures the quantity of the neighbors, while the significance considers also the "quality" of the neighbors. For example, if we order the nodes according to significance and according to degree we obtain two drastically different sequences. Here, we will limit ourselves to an indicative example. In graph Int-11-97, we have a node with degree 3 and significance 103.7, and a node of degree 10 and significance 1.305. The first node[4] connects to the three most significant nodes of the graph, while the second node does not connect to any node of high significance.

Significant nodes tend to be in the center of the network. The significance and the effective eccentricity are correlated. In figure 2,

[3]Note that we are focusing on the *topological* importance of a node, which is not necessarily related to other types of importance such as financial, or functional (amount of traffic that goes through a node).

[4]Note that significance here is according to our definition, and captures the topological significance and not the role of the node in the forwarding of traffic.

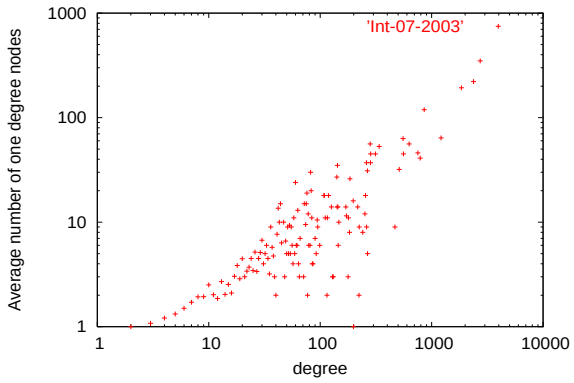

Figure 3: The average number of one-degree neighbors versus the degree of such node.

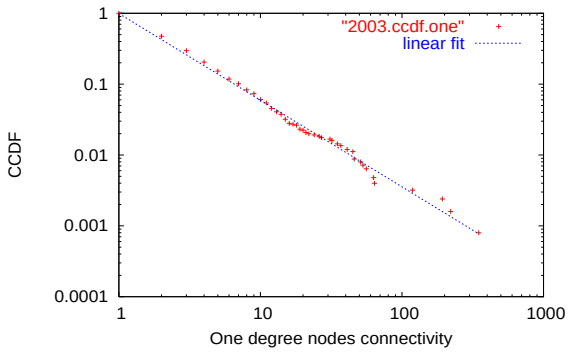

Figure 4: The CCDF of the one-degree neighbors of a node in log-log scale.

we plot the logarithm of the significance versus the effective eccentricity. We observe that nodes of high significance tend to have low effective eccentricity.Intuitively, this can be seen in two ways: central nodes are also significant, or that significant nodes gravitate towards the center.

The effective eccentricity of adjacent nodes cannot differ by more than one.

Lemma 1 *Let G=(V, E) be a connected undirected graph and (u, v) an edge in E, then the effective eccentricity of nodes u and v can not differ by more than one:*

$$|ecc(u) - ecc(v)| \leq 1$$

This lemma is easy to prove, since for any node x that node u can reach in h hops, node v can reach it with at most $h + 1$ hops.

This lemma helps us interpret the difference between the eccentricity of adjacent nodes. We can estimate the position of adjacent nodes with respect to the center of the network. When does this maximum difference in eccentricity appear? It does, when all paths are passing through a node. For example, consider a node of degree one: its eccentricity is equal to the eccentricity of its single neighbor plus one. We will refer to this observation when we evaluate the model we develop.

3.1 Location of One-Degree Nodes

The most common way to picture a hierarchy is to think of a social or military structure, where each class member connects to nodes of comparable importance. Each class connects to an immediately higher and lower class. Please refer to appendix B for a more detailed discussion on this model, which we call the **cast or broom** model. It turns out that the Internet topology deviates significantly from such a hierarchical model.

One-degree nodes are scattered all over the network. We examine the spatial distribution of the one-degree nodes in the graph. Note that one-degree nodes, are approximately $35 - 45\%$ of the nodes. In figure 3, we plot the average number of one degree nodes, that are adjacent to a node, versus it's degree. The qualitative observation is that one-degree nodes connect to both high and low degree nodes. Namely, the connectivity is not selective on the degree: nodes of the lowest degree can connect directly to the top nodes.

To better understand and characterize the one degree nodes, we examine the distribution of the one degree neighbors of a node. We use the Complementary Cumulative Distribution Function (CCDF) of the one degree neighbors of a node, which we denote as O_r. We plot the O_r versus the number of one degree neighbors r in log-log scale in figure 4 for graph Int-07-2003.

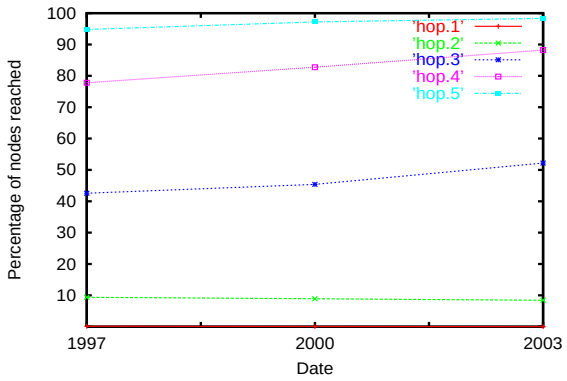

Figure 5: Percentage of nodes reached versus date for the topological paths. Each line represents the percentage for different number of hops.

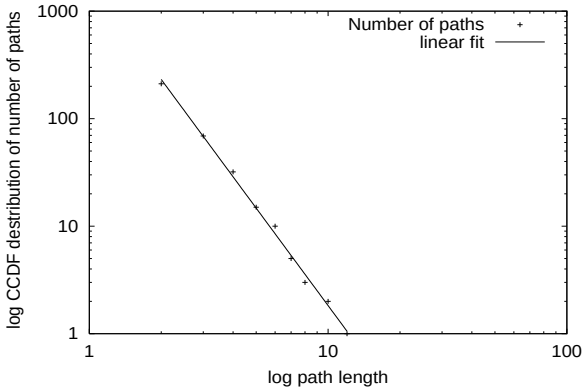

Figure 6: The CCDF distribution of the paths versus the path length between the nodes with degree 590 and 524 in log-log scale (Int-11-97). Correlation coefficient 99.8%.

The correlation coefficient is 99.5% for this instance and above 98% for all the instances we examined. This observation can be stated as the following power-law.

Power Law 4: Given a graph, the CCDF O_r of the one degree neighbors r of a node, is proportional to the r to the power of a constant θ.

$$O_r \propto r^{\theta}$$

A natural question to ask, is whether this power-law relates to the power-law of the degree distribution. Although there is a correlation between the degree of a node and the number of its one-degree neighbors, we do not observe a straightforward relationship such as a proportionality.

3.2 The Network Connectivity

To further examine the structure, we quantify the connectivity with two complementary metrics: a) the topological distances of the graph, b) the number of alternative paths that exist between two nodes.

Topological distances. The distribution of the topological distances has remained practically the same. We find that the distances in the network do not change significantly in the time period we examine. This is somewhat counter-intuitive for every-day thinking, where we expect something that increases in size to increase in all its dimensions.

In Figure 5, we plot the percentage of nodes we can reach for a given number of hops versus the day that each instance was collected. Each line corresponds to a different number of hops. We see that the neighborhood of a node as a percentage of the total nodes is either constant or increasing. For example, we find that within five hops we can always reach more than 95% of the nodes, and at least 45% of the nodes are within three hops. Given that the network increases, it is clear that the size of the neighborhood in absolute size is increasing for all the hops.

Properties of Alternate Paths. We study the number of alternate paths between a given pair of nodes, which provides a different aspect of the connectivity. There are several ways of defining alternate paths depending on the intended use, such as node-disjoint, edge-disjoint, or shortest paths only. Here, we are mostly interested in the topological insight we can obtain from the analysis.

Greedy shortest-path discovery method. We decided to emulate the behavior of a network operation, namely that of a possible fault-tolerant

routing protocol. Such a protocol may select the shortest path as primary path, and the second shortest path as back up. We restrict the back up path to be node disjoint with the primary. Following this, in our study, for each pair of nodes, we iteratively find and remove the shortest path, except the end points. We stop when we cannot find any more paths. Note that we have also considered another approach based on the max-flow method, where we maximize the number of alternate node-disjoint paths, but the results were comparable and are not shown here.

We find that the relationship between the number of node-disjoint paths and path length between a pair of nodes u and v is skewed. In figure 6, we plot the relationship between the CCDF distributions of the number of paths versus the path length for a pair of nodes. We want to capture this skewed path distribution concisely in a qualitative way. This leads us to state the following power-law as a rough approximation of the above observations. Recall that our focus is the topological structure and not an accurate model for the path length distribution.

Approximation Power Law 5: The complementary cumulative distribution function of the number of paths $R_{u,v}$ of length $l_{u,v}$ between a pair of nodes u and v (found by our greedy shortest path discovery method) is inversely proportional to the length of that path $l_{u,v}$ to the power of a constant m.

$$R_{u,v} \propto l_{u,v}^{-m}$$

The failure of the doughnut model. The value of the above observation is its insight on the macroscopic structure of the topology. Let us assume that the Internet topology is like a roughly homogeneous "doughnut", which we discuss further in appendix B. In this model, for any two nodes we would have two equally popular path lengths, each corresponding to one side of the doughnut. In that case, the path length distribution would not follow the power-law we observe in practice.

4 The Jellyfish Model

In this section, we integrate all the observations and insight of the previous sections into a conceptual model. First, we identify a topological center and we classify nodes into layers with respect to the center. Then, we show how the jellyfish model captures all the properties that we examined.

4.1 Core, Layers and Hierarchy

The first step in defining a hierarchy is to identify a starting point. A natural point to look for a center is the most important node. In fact, we observe that the highest degree nodes are adjacent to each other. We define the **core** as a **clique of high-degree nodes** with the following procedure. We sort nodes in non-increasing degree order. We select the highest degree node as the first member of the core. Then, we examine each node in that order; a node is added to the core only if it forms a clique with the nodes already in the core. In other words, the new node must connect to all the nodes already in the core. We stop when we can not add any more nodes. This way, the core is a clique but not necessarily the maximal clique of the graph.

Why do we define the core as a clique? Intuitively, the clique makes the representation more useful when we consider node distances, and in particular we can prove easily upper bounds for the distances of two nodes, as we discuss later in section 4.2. A path passing from the clique will have at most one hop through the clique. Thus, the distance between two nodes is bounded from their relative from the clique plus the one hop in the clique. In appendix A, we explore alternative definitions for the core by relaxing the stringent clique requirement. Recent, a work by Bar et al. examines the definition of an Internet core [33].

We now classify the rest of the nodes according to their proximity to the core. We define the first **layer** to be all the nodes adjacent to the core. Similarly, we define the second layer as the non-labeled neighbors of the first layer. By repeating

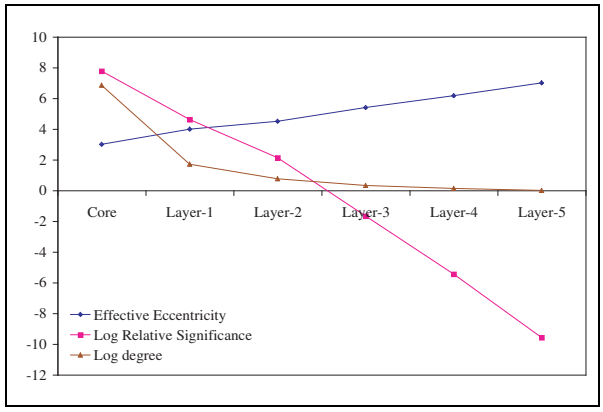

Figure 7: The average importance of each layer: log of the average degree, average effective eccentricity and log of the average relative significance (Int-07-2003 instance)

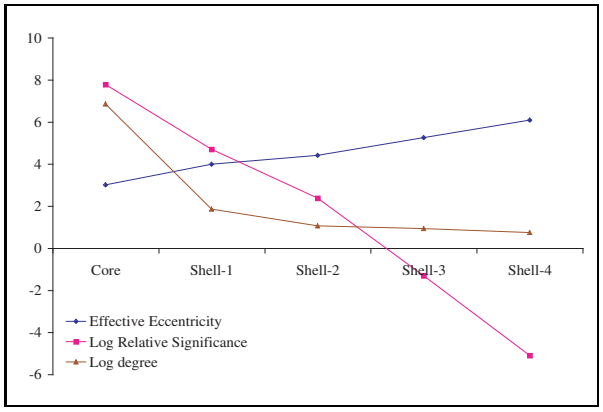

Figure 8: The average importance of each shell: log of the average degree, average effective eccentricity and log of the relative significance (Int-07-2003.

this procedure, we identify six layers if we count the core as a **layer** zero.

Node distribution across layers. Table 1 shows the distribution of the nodes for three Internet instances. The node distribution across layers does not seem to change significantly in the instances we examine. We make two interesting observations:

- Approximately 80-90% of the nodes are in the first 3 layers.
- We find six layers in all the instances we examine, and despite the significant network growth.

These two observations strongly suggest that the network grows "horizontally" by populating its layers and not by adding more layers. We elaborate on this point in the next section.

The effectiveness of the classification. We want to examine the effectiveness of our classification and explore its topological meaning. First, we find that the layers differ significantly in topological importance, which indicates that the classification captures some elements of the topological structure. We use our three metrics to quantify the importance of each layer. Figure 7 shows the average values of the effective eccentricity, the logarithm of the degree, and the logarithm of the relative significance for each layers for the Int-07-2003 instance. All metrics suggest that the importance of the nodes of each layer decreases rapidly as we move away from the core. Note that for the average degree and the relative significance the scale is logarithmic.

It is important to locate and study separately the one-degree nodes. First, the one-degree nodes are not useful in terms of connectivity to the rest of the network. Second, one-degree nodes are a large percentage of the network, and it is important to clarify and isolate their role. We separate each layer into two classes: a) the multiple-degree or **shell** nodes, and b) the one-degree or **hang** nodes. We refer to the one-degree nodes hanging from k-th shell as the k-th hang class. The layers and the shells have the following relationship:

$$Layer_k = Shell_k + Hang_{k-1}$$

For example, shell-0 is the core, and its one-degree neighbors are denoted as hang-0, while the rest of the neighbors constitute shell-1.

Table 2 shows the size of each group of nodes in our classification.

The topological importance of shell decreases as we move away from the core. In figure 8, we plot the logarithm of the average de-

	Instance					
	Int-11-1997		Int-06-2000		Int-07-2003	
Layer No	Nodes	% of Nodes	Nodes	% of Nodes	Nodes	% of Nodes
Core/Layer-0	8	0.23	14	0.176	13	0.08
Layer-1	1354	44.90	3659	46.25	7330	46.27
Layer-2	1202	39.866	3090	39.05	7116	45.51
Layer-3	396	13.134	1052	13.29	1078	6.89
Layer-4	43	1.425	86	10.87	96	0.61
Layer-5	12	0.398	10	0.12	1	0.0063

Table 1: Distribution of nodes in layers for three Internet instances.

	Instance					
	Int-11-1997		Int-06-2000		Int-07-2003	
Layer ID	Nodes	% of Nodes	Nodes	% of Nodes	Nodes	% of Nodes
Core/Shell-0	8	0.23	14	0.176	13	0.08
Hang-0	465	15.42	798	10.08	1174	7.5
Shell-1	889	29.49	2861	36.16	6156	39.37
Hang-1	623	20.66	1266	16	2821	18.04
Shell-2	579	19.2	1824	23.05	4295	27.47
Hang-2	299	9.92	662	8.36	808	5.16
Shell-3	97	3.22	390	4.92	270	1.72
Hang-3	41	1.36	74	0.93	84	0.53
Shell-4	2	0.66	12	0.15	12	0.07
Hang-4	12	0.4	10	0.12	1	0.006

Table 2: Distribution of nodes in shell and hang classes.

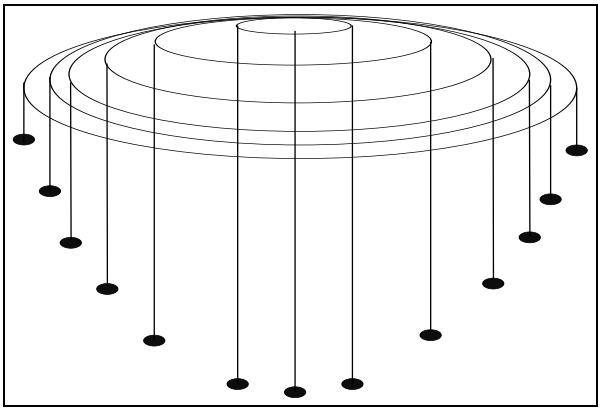

Figure 9: The Internet topology as a jellyfish.

gree distribution, the logarithm of average relative significance, and the average effective eccentricity of each shell. All metrics suggest that the importance of shells near the core is higher. Note that for the average degree distribution and significance the scale is logarithmic so a difference of one is substantial. This analysis indicates that our shells manage to cluster the nodes according to their topological importance.

Most of the connectivity is towards the center. Observe that the average effective eccentricity increases by approximately 0.5 to 1 as we go away from the core. Recalling the lemma of section 3, an increase in effective eccentricity of approximately one indicates that the outer node is approximately one link further away from the core. Intuitively, nodes at the outer shells need to go through the previous shell for most of their shortest path connections. This suggests that our selection of the core and the layers captures effectively the direction of the paths.

4.2 The Jellyfish Model

We use the shell-hang classification to define the jellyfish model. The core is the center of the head of the jellyfish surrounded by shells of nodes. Figure 9 shows a graphical illustration of this model. The hang nodes form the tentacles of the jellyfish. We make the length of the tentacle longer to graphically represent the concentration of one-degree neighbors for each shell. We can color each shell according to its importance.

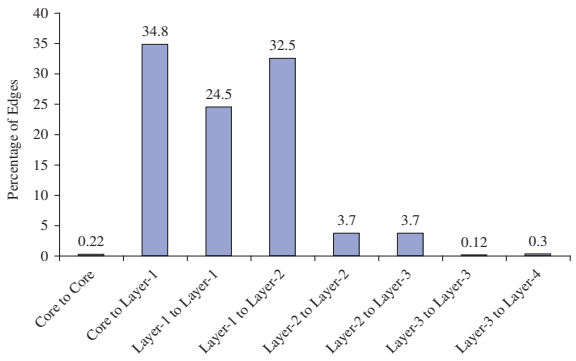

Figure 10: The percentage of the different types of edges classified according to their adjacent nodes (Int-07-2003).

From table 2 we can see how the nodes are distributed in the different layers of our model. We observe that $80 - 90\%$ of nodes are in the first few layers: core, shells one and two, and hang one and two.

The jellyfish model provides upper bounds of the node distances. By construction, the model can provide upper bounds on the distances between nodes. For example, we can state that **the nodes in the first three layers (80-90% of nodes) are within 5 hops from each other**[5]. In the worst case, the shortest path between two nodes in layer two would consist of nodes in: layer-2, layer-1, core, core, layer-1, layer-2. The distance is bounded by five, but it could be less than five. Note that this upper bound seems to be very close to empirical observations, that indicate that 90% of the nodes are within 5 hops [34]. Generalizing, we can prove the following lemma.

Lemma 2 (Upper Bound of Distance): *The distance between nodes v of layer-k_v and w of layer-k_w is bounded above as follows:*

$$d(v, w) \leq k_v + k_w + 1$$

[5]We refer to layers instead of the equivalent shell and hang for simplicity.

The proof is a straightforward if we consider the construction of the layers.

In the jellyfish model, 70% of edges are between different node layers. In Figure 10 we plot the percentage of edges that exist between and within layers. We find that 70% of the edges connect nodes between different layers. We think of these edges as **v**ertical with respect to the jellyfish hierarchy. In contrast, approximately 30% are horizontal to the hierarchy providing connectivity between nodes of the same class.

Let us examine the vertical edges in more detail. The construction of the jellyfish model is a breadth-first type of network exploration. The breadth-first tree consists of $N-1$ edges, where N the number of nodes. The number of edges in the graph is approximately: $2N$ (average degree close to four). Therefore, 50% of the edges are part of the breadth first tree of the jellyfish. Recall that 70% of the edges are vertical edges. This means that the other $70-50=20\%$ of the vertical edges are "redundant" edges.

Why is the jellyfish a good model? It should be clear by now that this model is driven by several empirical observations. We provide an overview of the topological properties that the jellyfish model captures. As an intuitive model, the jellyfish represents these properties in a graphical and qualitative way[6].

1. **Core:** The topology has a core of highly connected topologically-important nodes, which is represented by the center of the jellyfish cap.

2. **Five layers:** The distances between nodes are small; maximum distance less than 11 hops, and 80% of nodes are within 5 hops.

3. **Center-heavy:** 80% of the nodes are in the first 3 layers.

[6]Note that not all properties listed below can be deduced directly from the model, but the model can act as a intuitive reminder.

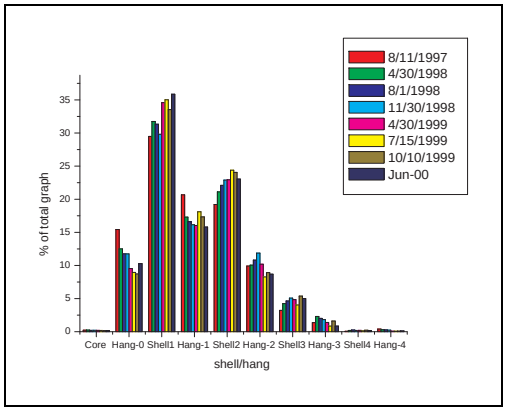

Figure 11: The distribution of nodes over time grouped by class.

4. **One-degree nodes:** There is a non-trivial percentage (35-45%) of one-degree nodes, which are scattered everywhere (representation of power-law 4).

5. The importance of the nodes decreases with their distance from the core.

6. The model provides good upper bounds of the distances between nodes.

An additional strength of the model is that it has persisted in time. Its structure and the node distribution across classes has not changed qualitatively during the years of our study. We elaborate on the model evolution in the next section.

5 The Evolution of the Jellyfish

We study the evolution of the Internet structure for approximately three years. We find that the statistical properties of the jellyfish model change relatively little over time. Second, we find that the node and edge distribution across the jellyfish classes remains approximately the same. Third, we find that the classification of individual nodes does not change significantly within a three to eight month interval. Finally, we study the identity of the nodes that constitute the core.

	Nov97 v/s Apr98	Apr98 v/s Aug98	Nov98 v/s Apr99	Apr99 v/s July99	July99 v/s Oct99	Oct99 v/s June00
No change	1845	2829	3278	3986	4289	3972
Total change	968	651	887	919	963	1615
Hang to shell	343	228	395	312	320	634
Shell to hang	148	155	227	193	265	382
Drop in shell	119	69	97	152	54	256
Increase in shell	117	97	60	62	171	140
Drop in hang	91	38	67	155	27	125
Increase in hang	150	64	41	45	126	78

Table 3: The change in the node classification between consecutive instances.

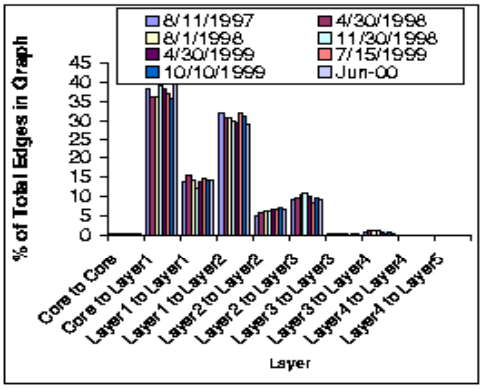

Figure 12: The distribution of edges over time grouped by type.

Distribution of nodes over time: horizontal growth. The first striking observation is that the growth does not create new layers: **the network grows horizontally**. The second observation is that the distribution of the nodes in each class remains approximately the same, see figure 11. The percentage of nodes in each category is within ±5% with respect to the average of the category[7] over the 8 instances and figure 12 shows the average node distributions over time.

Changes in the classification of individual nodes. The classification of most of the nodes in the jellyfish model does not change significantly within a three to eight month interval. Table 3 shows the results of comparing multiple pairs of instances. We list the number of nodes that have remained in the same class, or changed classes. In case of change, we have several different types depending on whether the move was from a shell to a hang category or between different shells etc. Naturally, we consider only nodes that appear in both instances under comparison. We find that most of the nodes (about 60% to 80%) maintain the same classification. The second largest type of change is nodes going from a hang to a shell category. This suggests that the existing nodes become more connected with time possibly to increase their fault-tolerance, which has been an independently observed trend [16].

Edge distribution over time. We look at the distribution of edges over time with respect to our model[8]. The main observation is that the edge distribution among the layers stays within ±5% of the average of each layer. Figure 12 shows the edge distribution over time. Observe that the maximum number of edges are between core and layer-1 and between layer-1 and layer-2. These two groups account for about 65% of all the edges in the graph.

Which ASes are in the core? We find that there are 20 nodes that appear in the core at

[7]Note that the 5% refers to the total number of nodes, and not 5% of the average of the class. With this definition, a change from 30% to 35% is within 5%.

[8]We refer to layers instead of shell-hang classes to simplify the plot.

least once in the eight instances we examine. Among them, only four nodes appear in all eight instances. The maximum number of nodes in the core in any instance was 14 and the minimum was 8. Here are the most interesting observations:

- The maximum number of nodes in the core is 14 in June 2000, while the minimum is 8 in November 1997 and August 1998.
- There were four nodes that are always in the core: AlterNet, Cable and Wireless, Sprint-Link and GTE Internetworking. These nodes also constitute the highest-degree nodes in the core, except in June 2000 when AT&T (572) exceeded GTE (426)
- The node with the smallest degree in the core was Exodus communications (53) in April 1998.

6 The Importance of the Jellyfish Model

So far, we showed that the jellyfish can be used to describe the Internet in a consistent way for the last six years. In this section, we demonstrate the usefulness of jellyfish. There are two parameters that we need to investigate. First, we try to answer whether all graphs can be described using the jellyfish model that we found in the previous sections. Ideally, jellyfish should be able to distinguish among different type of graphs. We try to answer this in a qualitative way by using simple regular topologies, and we show that not all topologies can be characterized as jellyfish. Second, we check whether our model can be used to distinguish among power-law graphs[9]. Using two popular graph generators, we show that graphs that follow approximately the same power-law distribution, can have significant different macroscopic properties, and can be distinguished using jellyfish.

[9]By power-law graphs, we mean graphs that their degree distribution follows a power-law

Can any graph be modeled as a jellyfish? For every graph, we can pick a center and compute it's layers. On the other hand, not every graph can match the Internet profile, i.e. the number of layers and the distribution of nodes among these layers. For example, let us consider some regular topologies such as a square mesh, a complete binary (or k-ary) tree, a clique [10], and purely random graphs. None of these graphs will fit the above description of the jellyfish in all its aspects. As an example, we will mention a few of the more pronounced differences. First, there is no natural central point to place the core. Even if we define an arbitrary core, there are also other properties that will be violated. The mesh and the tree will have a large number of shells proportional to $O(\sqrt{N})$ or $O(\log N)$ respectively. More importantly, when the size of the network would double, the number of layers and shells would increase, which does not happen here. The clique also does not fit the jellyfish profile: it has only one layer, and no one-degree nodes. We have also seen other network models which do not match this profile such as the strictly hierarchical model, and the doughnut model in appendix B.

6.1 Power-law Graph Generators

We can use the Jellyfish model as a test of the realism of Internet like graphs. We will use the GLP methodology proposed in [6], and the PLRG approach proposed in [1]. The GLP approach depends on the preferential model, and it is the most recent proposed generator and is considered to be the state of the art. On the other hand, the PLRG generator is based on an interesting theoretical model for scale free graphs, and takes the degree distribution as a given. Note that in [6], they compared the two generators and found that the best generator is the GLP. They showed that PLRG fails to capture properties like the characteristic path length and the clustering coefficient. We show that GLP

[10]Varying these models by adding or removing a few edges or nodes in a uniformly distributed way will not reconciliate the differences with the jellyfish in most cases.

	Instance					
	GLP		Int-06-2000		PLRG	
Layer ID	Nodes	% of Nodes	Nodes	% of Nodes	Nodes	% of Nodes
Core/Shell-0	21	0.2	14	0.176	11	0.13
Hang-0	1885	23.82	798	10.08	565	7.1
Shell-1	1672	21.13	2861	36.16	2346	29.6
Hang-1	3371	42.6	1266	16	1298	16.4
Shell-2	688	8.7	1824	23.05	2305	29.13
Hang-2	221	2.79	662	8.36	525	6.6
Shell-3	3	0.037	390	4.92	325	4.1
Hang-3	3	0.037	74	0.93	125	1.5
Shell-4	0	0	12	0.15	41	0.51
Hang-4	0	0	10	0.12	23	0.29

Table 4: Distribution of nodes in shell and hang classes.

does not capture the macro structure that we found using jellyfish. Incidentally, PLRG seems to pass the test, although it fails other properties. Therefore, our jellyfish model is an excellent tool to distinguish graphs.

We use the Brite generator [23] for the GLP model, which includes an implementation of this model [11]. We have implemented PLRG. In order to compare the Internet topology with the generators, we will use the Int-06-2000 graph. We want to generate a topology that would have the same properties as Int-06-2000. Following the methodology presented in [6] we use the following parameters: $\rho = 0.434$ and $\beta = 0.661$ for the GLP. For the PLRG, we simply use the degree distribution of Int-06-2000.

The correlation coefficient for the degree power-law plot is $97,6\%$ for the GLP with slope $a = -1.092$. For the PLRG plot the correlation coefficient is $99,7\%$ and the slope is $a = -1.243$. For the Int-06-2000 the correlation coefficient is 99.7% and the slope is $a = -1.163$[12]. In table 4, we have the decomposition of the graphs using the jellyfish model. These results clearly show that the generated graph using the GLP methodology is qualitatively different than the Internet graph. On the other hand PLRG seems to maintain similar structure according to the jellyfish model. The only differences between PLRG and Int-06-2000 is that the clique is smaller, having only 11 nodes, and that we have a slightly smaller shell-1 and bigger shell-2.

Where does GLP fail to model the Internet? The main differences between GLP and the Int-06-2000 can be summarized as following:

1. The core of the network is much bigger in GLP compared to the Internet. More specifically, we have 21 nodes in the core for the GLP, while only 14 nodes in the Int-06-2000.

2. The number of hanging nodes (degree one) far out-exceeds the number of shell nodes. The analogy is approximately 70% hanging nodes to 30% shell nodes. In the case of the Int-06-2000 we have the opposite result.

3. The GLP topology has only up to 5 layers, with the 5th layer having only 3 members,

[11] Note that the GLP model in the Brite generator is not exactly the same as described in the original paper. The difference lies in that the number of edges of a new node can be either one with probability 87%, or two with probability 13%. We updated the model used in brite to reflect the original approach.

[12] Note that the PLRG doesn't have the exact same degree distribution as the Int-06-2000. When we generate the topology using PLRG, we might pick to connect two nodes that are already connected, so in this sense we have fewer edges in the final graph.

while the Int-06-2000 has 6 layers.

Using our analysis we can conclude that jellyfish is an excellent tool to distinguish among generators into two classes, those that can capture the macro structure of the Internet and those that can not.

7 Conclusions

In this paper, we develop a simple and conceptual topological model for the inter-domain Internet topology. Our work has five components of independent interest. First, we present and study three metrics of the topological importance of a node. Second, we identify some new topological properties. Third, we integrate the main properties into our jellyfish model. Fourth, we study the time evolution of the Internet with respect to our model. Finally, we show that our model can be used to distinguish among graph generators.

The jellyfish model provides novel insight into the structure of the Internet topology. Despite its simple nature, the jellyfish captures most of the known macroscopic properties. The model facilitates the visualization of the complex Internet structure by abstracting it into something that a human can easily picture and understand.

We summarize our main observations and contributions in the following points.

- We use three metrics to quantify the topological importance of a node and we examine their meaning and their relationships.
- The Internet has a highly connected core and layers of nodes in decreasing importance. This way, we can define a notion of loose hierarchy in the network.
- The jellyfish model provides fairly tight upper bounds of the distances between nodes.
- Low degree nodes are scattered in the network in contrast to a strictly layered hierarchy.
- Approximately 30% of the edges are between nodes of the same class according to our model. From the remaining edges, 20% of the edges are "redundant" edges between adjacent layers.
- The topological growth is horizontal: the number of layers has not increased over time.
- The statistical properties of the topology with respect to the jellyfish have not changed significantly over time.
- The jellyfish can be used to distinguish among graph generators.

In the future, we want to develop a theoretical framework that will explain and justify our experimentally derived model. The ground breaking work of Reittu and Norros [32] opens the doors for a parallel approach where theory and real-data analysis complement each other. Furthermore, we intend to elaborate and fine tune the jellyfish model by integrating more topological properties. We want to identify more topological properties and integrate them into the model using novel means such as color. Finally, we want to examine whether other real networks can be described by the jellyfish model.

References

[1] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. *STOC*, pages 171–180, 2000.

[2] R. Albert and A. Barabasi. Topology of complex networks:local events and universality. *Phys.Review Letters*, 85, 5234, 2000.

[3] R. Albert, H. Jeong, and A. Barabasi. Attack and error tolerance of complex networks. *Nature*, 406, 378, July 2000.

[4] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286, 509-512, October 1999.

[5] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and

J. Wiener. Graph structure in the web: experiments and models. *In Proc. of the 9th World Wide Web Conference*, 2000.

[6] T. Bu and D. Towsley. On distinguishing between Internet power law topology generators. *IEEE Infocom*, 2002.

[7] Qian Chen, Hyunseok Chang, Ramesh Govindan, Sugih Jamin, Scott J. Shenker, and Walter Willinger. The origin of power laws in Internet topologies revisited. *IEEE Infocom*, 2001.

[8] Bill Cheswick and Hal Burch. Internet mapping project. Wired Magazine, December 1998. See http://cm.bell-labs.com/cm/cs/who/ches/map/index.html.

[9] Reuven Cohen and Shlomo Havlin. Scale-free networks are ultrasmall. *Physical Letters Review*, 90(5), 2002.

[10] F. Comellas, G. Fertin, and A. Raspaud. Vertex labeling and routing in recursive clique-trees, a new family of small-world scale-free graphs. *Sirocco*, 2003.

[11] M. Doar. A better model for generating test networks. *IEEE Global Internet*, Nov. 1996.

[12] A. Fabrikant, E. Koutsoupias, and C.H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the Internet. *ICALP*, Springer-Verlag LNCS:110–122, 2002.

[13] Lixin Gao. On inferring automonous system relationships in the Internet. *IEEE/ACM Transactions on Networking*, 9:733–745, December 2001.

[14] Z. Ge, D. Figueiredo, S. Jaiswal, and L. Gao. On the hierarchical structure of the logical internet graph. *SPIE ITComm*, 2001.

[15] R. Govindan and A. Reddy. An analysis of Internet Inter-domain topology and route stability. *IEEE Infocom*, Kobe, Japan, April 7-11 1997.

[16] Geoff Huston. Homepage:. http://www.telstra.net/gih, 2001.

[17] Young Hyun. Walrus project. *http://mappa.mundi.net/maps/maps_020/*, 2002.

[18] Cheng Jin, Qian Chen, and Sugih Jamin. Inet: Internet topology generator. *Techical Report UM CSE-TR-433-00*, 2000.

[19] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 1999. (Earlier version in ACM-SIAM Symposium on Discrete Algorithms, 1998).

[20] L.Subramanian, S.Agarwal, J.Rexford, and R.Katz. Characterizing the Internet hierarchy from multiple vantage points. *IEEE Infocom*, 2002.

[21] Damien Magoni and Jean Jacques Pansiot. Analysis of the autonomous system network topology. *ACM Sigcomm Computer Communication Review*, 31(3):26–37, July 2001.

[22] A. Medina, I. Matta, and J. Byers. On the origin of powerlaws in Internet topologies. *ACM Sigcomm Computer Communication Review*, 30(2):18–34, April 2000.

[23] Alberto Medina, Anukool Lakhina, Ibrahim Matta, and John Byers. BRITE: Topology generator. *http://www.cs.bu.edu/brite/*.

[24] M.Mihail and C.H.Papadimitriou. On the eigenvalue power law. *Random*, 2002.

[25] Jae Dong Noh. Exact scaling properties of a hierarchical network model. *Physical Review E*, 67(045103), 2003.

[26] University of Oregon Route Views Project. Online data and reports. http://www.routeviews.org/.

[27] CAIDA Org. Skitter project. *http://www.caida.org/tools/measurement/skitter/*, 2002.

[28] C. Palmer, G. Siganos, M. Faloutsos, C. Faloutsos, and P. Gibbons. The connectivity and fault-tolerance of the Internet topology. *Workshop on Network-Related Data Management (NRDM 2001), In cooperation with ACM SIGMOD/PODS, Santa Barbara*, 2001.

[29] J.-J. Pansiot and D Grad. On routes and multicast trees in the Internet. *ACM Sigcomm Computer Communication Review*, 28(1):41–50, January 1998.

[30] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 2nd edition, 1992.

[31] Erzsebet Ravasz and Albert-Laszlo Barabasi. Hierarchical organization in complex networks. *Physical Review E*, 67(026112), 2003.

[32] H. Reittu and I. Norros. On the power law random graph model of the internet. *Tech. Report, VTT Information Technology*, 2002.

[33] M. Gonen S. Bar and A. Wool. An incremental super-linear preferential internet topology model. *5th Annual Passive and Active Measurement Workshop (PAM)*, 2004.

[34] G. Siganos, M. Faloutsos, P. Faloutsos, and C. Faloutsos. Power-laws and the as-level Internet topology. *IEEE/ACM Trans. on Networking*, August 2003.

[35] H. Tangmurankit, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger. Network topology generators: Degree-based vs structural. *ACM Sigcomm*, 2002.

[36] L. Tauro, C. Palmer, G. Siganos, and M. Faloutsos. A simple conceptual model for the internet topology. *IEEE Global Internet*, November 2001.

[37] K. W. Koput W. W. Powell, D. R. White and J. Owen-Smith. Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American Journal of Sociology*, 110(4):1132–1205, 2005.

[38] B. M. Waxman. Routing of multipoint connections. *IEEE Journal of Selected Areas in Communications*, pages 1617–1622, 1988.

[39] D. R. White, J. Owen-Smith, J. Moody, and W. W. Powell. Network dynamics and field evolution: The growth ofinterorganizational collaboration in the life sciences. *Computational and Mathematical Organization Theory*, 10(1):95–117, 2004.

[40] E. W. Zegura, K. L. Calvert, and M. J. Donahoo. A quantitative comparison of graph-based models for internetworks. *IEEE/ACM Transactions on Networking*, 5(6):770–783, December 1997. http://www.cc.gatech.edu/projects/gtitm/.

A Core: Relaxing the clique constraint

We explain our choice of defining our core to be a clique containing the maximal degree node. One could consider near-cliques, and include in the core nodes of high degree that connect with almost all the core nodes. This would make sense if by relaxing the clique constraint, we could get a set of nodes that are substantially more important than the nodes left out of the clique.

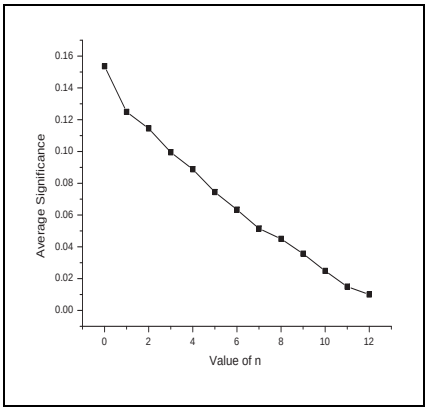

Figure 13: Significance of the core versus relaxing the clique constraint by n edges.

Definition 1: We define an n-relaxed clique to be a set of nodes that connect to every node in the set except at most n nodes.

For example, a set of k nodes is a 1-relaxed clique if and only if each node connects to at least d-1 nodes in the original clique where d is the size of the original clique.

We do the following experiment. We begin with the clique containing the highest degree node. Let c be the number of nodes in the clique. This corresponds to a 0-relaxed clique. Now for the next iteration, we relax this requirement and allow nodes into the core that have one missing edge from the original core forming a 1-relaxed clique.

We plot the average significance of an n-relaxed clique versus n for Internet instance Int-06-2000. We vary n from 0 to $c-1$ where c is the number of nodes in the first clique ($n = 0$).

First, we observe that the maximum average significance is obtained when the core is a clique which corresponds to the case of $n = 0$. Second, we see that as we increase n the average significance of the core decreases smoothly. Therefore, we do not have a reason to pick a value of n other than zero. For n=0, we have the maximum average significance and also the interpretation of

Property	Cast	Furball	Doughnut	Jellyfish
Horizontal Connectivity	no	yes	yes	yes
High-Low Degree Edges	no	no	maybe	yes
Distance Distribution	yes	yes	no	yes

Table 5: The matrix of observed properties and wether they are satisfied by the different models.

the core is straightforward.

B Failed Internet Models

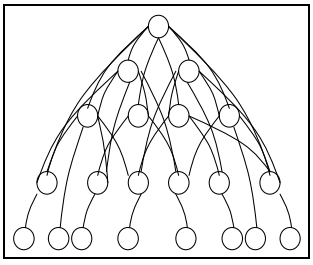

Figure 14: The Internet topology as a broom.

In this section we take a look at several models that looked promising but failed, as they could not model the properties that we observed (see table 5).

The Cast or Broom Model. This model is probably the simplest model one could visualize where domains are connected as parent and children. However this model fails because we do not take into account that ASes could be connected **horizontally** as peers and it does not capture that one-degree nodes connect to high degree nodes as we explain below.

The Furball Model. This model allows for nodes to be classified into layers as before. However it assumes a connectivity scheme via which the high degree nodes only connect to other high degree nodes and so on with the one-degree domains connecting to the edge of the network. However this violates our power-law on the distribution of one-degree nodes which states that the one-degree nodes are uniformly distributed throughout the network and thus, the model fails.

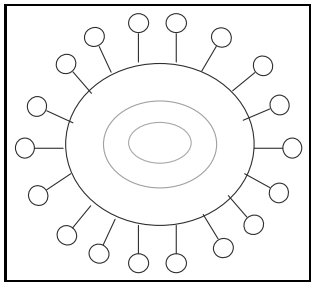

Figure 15: The Internet topology as a furball.

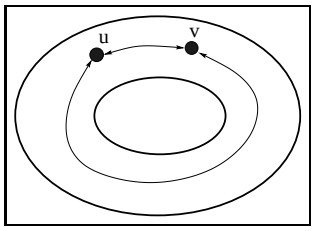

Figure 16: The Internet topology as a doughnut.

The Doughnut Model. Here we try to model the Internet as a ring. The figure shows the possible paths between two nodes in a layer. In this model there are several paths of short length between any two nodes. However there are also several paths that go all the way around the previous layer. This model fails as we know that the majority of nodes go through the core to connect to other nodes. Therefore we do not find long round paths as proposed by this model.