

# Characterizing and Modelling Clustering Features in AS-Level Internet Topology

Yan Li, Jun-Hong Cui, Dario Maggiorini and Michalis Faloutsos

UCONN CSE Technical Report: UbiNet-TR07-02

Last Update: July 2007

## Abstract

The AS-level Internet topology has shown significant clustering features. In this paper, we propose a new set of clustering metrics and conduct extensive measurement on the AS-level Internet topology. We give a thorough characterization on the clustering features and their evolution. We also study the clustering features of different topological structures by comparing the Internet with various topology models. Due to the limitation of existing topology models on capturing clustering features, we design a new topology model based on clustering. Through extensive evaluations, we claim that our model can closely capture the clustering features as well as other common topological properties.

## I. INTRODUCTION

Recent work has shown that the AS-level Internet topology has significant clustering features [5], [12], [18], [19]: In the abstracted graph, some nodes are densely connected and form clusters which are sparsely interconnected. Accurately characterizing and modelling these clustering features is very important in various strands of networking researches. First grasping the clustering features can make significant improvement in network modelling and simulations [5], [6], [12], [18], [19], [21]. Also, the clustering features of the AS-level Internet topology have direct impact on the performance of inter-domain routing protocols. Clustering ASes would allow more efficient inter-domain routings [9], [11], [14]. Further, clustering features characterization can aid the analysis on network traffic, congestion, robustness and other critical network issues [18], [19].

In the literature, the major metric used to measure clustering features is *clustering coefficient*,  $C$ , which is originally defined as follows. Suppose that a vertex  $v$  has  $k_v$  neighbors; then at most  $k_v(k_v - 1)/2$  edges can exist between them. Let  $C_v$  denote the fraction of these allowable edges that actually exist. Then  $C$  is defined as the average of  $C_v$  overall all  $v$  [25]. From the definition, we can see that clustering coefficient provides a measure of how close a node and its neighbors are to forming a clique [19]. This is a very useful metric. For example, it to some extent expresses the local robustness in the graph, and thus has many practical implications [19]. However, clustering coefficient can only measure local clustering features, and it misses the capability to capture global clustering features. For example, for any node, how are its neighbors' neighbors (i.e., two-hop neighbors) connected? How about three-hop neighbors or more? How are clusters (i.e., some densely connected

nodes) correlated? How are clusters structured? To answer these questions, it is desirable to design a new set of metrics to accurately capture all levels of clustering features.

In this paper, we propose a novel set of metrics (based the concept of network clustering) for clustering features characterization. The proposed metrics include clustering significance, cluster size, and inter/intra-cluster connections. We conduct extensive measurement on the AS-level Internet topologies which demonstrate significant clustering features and a stable evolution trend. We also study the clustering features of various topologies generated from different topology models. By comparing these topologies with the Internet topology instances, we evaluate the capability of the existing synthetic topology models to reproduce the clustering features of the Internet. We observe that all existing topology models have limitations in matching those clustering metrics of the Internet. We thus design a new topology model, which is based on our proposed metrics. From the experiment results, we find that our new model performs better than previous topology models in capturing the clustering features of the AS-level Internet topology.

The rest of this paper is organized as follows. In Section II, we first briefly overview the concept of network clustering, clustering criteria and clustering algorithms. Then in Section III, we describe our methodology and propose a set of clustering metrics based on network clustering. After that, we show the clustering features of the Internet topology instances using the proposed metrics in Section IV and study the clustering features of different topologies generated by various topology models in Section V. In Section VI, we present our new topology model and show its power by simulations. Finally we discuss some related work in Section VII and conclude this paper in Section VIII.

## II. BACKGROUND

### A. Network Clustering

Clustering is an important technique used in biology, chemistry, linguistics, physics, sociology and variety of areas. Clustering can be loosely defined as a process of organizing objects into groups whose members share certain similarity. Network clustering is also called graph clustering since a graph is used to abstract a network topology in networking research. It provides a way to partition a network topology into clusters such that nodes in the same clusters are highly connected and between clusters they are sparsely connected.

In the literature, there has been considerable research effort in designing efficient network clustering algorithms [7], [23], [24], [16], [17]. This paper, however, is not on clustering algorithm design. Instead, we propose to use existing accurate network clustering algorithms to analyze AS-level Internet topology and study the clustering features.

### B. Clustering Criteria

Based on the definition of network clustering, node connectivity is the most important factor to be considered in a network clustering algorithm. A good clustering algorithm should guarantee nodes in the same clusters are

highly connected and between clusters are less connected. There are other issues to be addressed when applying clustering approaches into different applications, such as the cluster size, the number of orphan nodes, etc., but in our work, we aim to study the natural topology properties and should not bias on the clustering processes. Based on this criterion, we choose the Scaled Coverage Measure (SCM) [24] as the metric to quantify the accuracy of a clustering technique, and compare different clustering algorithms.

SCM is a practical clustering accuracy measure. The basic idea of SCM is in accurate clusterings, each node should be clustered only with its neighbors and the number of non-neighbor nodes in its cluster and neighbor nodes excluded from its cluster should be minimized. This idea exactly matches aforementioned criterion of good clusterings. Following is the definition of SCM.

Let  $V$  be the set of network nodes and  $E$  be the set of edges in a topology  $G$ . We denote a clustering on this topology as  $\mathcal{C} = \{C_1, C_2, \dots, C_l\}$  where  $\bigcup_{i=1}^l C_i = V$ . Under clustering  $\mathcal{C}$ , any particular node  $v_i$  is associated with the following sets:

**NBR**( $v_i$ ) is the set of neighbors of  $v_i$ ;

**CLUST**( $v_i, \mathcal{C}$ ) is the set of nodes that are in the same cluster as node  $v_i$  (Excl.  $v_i$ );

**NNBR**( $v_i, \mathcal{C}$ ) is the set of non-neighbor nodes which are in the same cluster as  $v_i$ .

**XNBR**( $v_i, \mathcal{C}$ ) is the set of neighbors of  $v_i$  but are excluded from **CLUST**( $v_i, \mathcal{C}$ ).

Then, the SCM value of node  $v_i$  is calculated by:

$$1 - \frac{|NNBR(v_i, \mathcal{C})| + |XNBR(v_i, \mathcal{C})|}{|NBR(v_i) \cup CLUST(v_i, \mathcal{C})|}. \quad (1)$$

The SCM value of the whole topology is simply calculated as the average of the SCM of all the nodes in the topology, and is bounded by 0 and 1.

SCM can well reflect the significance of the clustering features in a topology. The higher the SCM, the more significant the clustering features. For any graph, there exists a highest SCM value which is determined only by the topological structures and therefore is considered as a natural property of a graph. When comparing the accuracy of different clustering approaches on the same topology, the higher the SCM value, the more accurate the clustering approach.

### C. Clustering Algorithms

We evaluate various clustering algorithms DDP [7], MCL [24], CDC [23], and SACA [16] based on SCM, and select SACA as the clustering method for our clustering feature study, as it can achieve higher SCM values for various topologies compared to the other algorithms. More details about SACA can be found in [16].

## III. METHODOLOGY

Our clustering feature characterization procedure is as follows: 1) Obtain AS-level Internet topology instances; 2) Get accurate clusters using SACA; 3) Conduct measurements on the derived clusters. Next we present the Internet

topology data sources and the clustering metrics we propose.

### A. Topology Sources

Due to large scale of the Internet, it is very difficult if not impossible to get complete Internet topology snapshots, but partial views can be obtained from different data sources [19]. In this paper, the Internet topology instances we use are obtained from Oregon Routeviews [2] and CAIDA [1]. The Internet has been growing extremely fast [1]–[3]. Based on the topology instances collected from the Oregon Routeviews project, there are around 3000 ASes at the end of 1997, but in May 2007, the number of ASes has exceeded 25000. Over the 9 years, the AS-level Internet topology has been growing at a rate of more than 2000 ASes per year. This high growing speed results in a huge and complicated network which is very difficult to characterize. Fortunately, the evolution of the Internet is not purely random. In Section IV, we will measure the clustering features of AS-level Internet topologies, and examine how these features are well preserved.

### B. Clustering Metrics

We propose the following clustering metrics:

- **Clustering Significance** It is measured by the SCM values of topologies. Given a topology, the higher SCM value means more significant clustering features. Note that this metric is sensitive to both the structure and the scale of a topology.
- **Cluster Size** Cluster size is defined as the number of nodes in a cluster. Using regression techniques, we find the cluster size of AS-level Internet topologies follows certain distribution. This metric is highly dependent on the topology structures and therefore is considered as an important metric for clustering characterization.
- **Inter-cluster Connection** We say a pair of clusters are inter-connected if at least one pair of nodes in the two clusters are connected. For a given cluster, its inter-cluster connection is defined as the number of links between the nodes within this cluster and the nodes in the other clusters. The distribution of inter-cluster connection describes the peering relationships between clusters and reflects the hierarchical structure of a topology.
- **Intra-cluster Connection** This is the metric measuring the peering relationships between nodes within a cluster. For a specific node, the intra-cluster connection is defined as the number of links between this node and its neighbors within the cluster. The distribution of intra-cluster connection reflects the internal structure of clusters.

## IV. CLUSTERING FEATURES AND EVOLUTIONS

In this section, we study the AS-level Internet topology and its evolution in terms of clustering features.

We measure and examine the clustering features of the AS-level Internet topology snapshots over a long time period. The topologies presented in this paper are obtained from Routeviews' BGP and CAIDA's Skitter traceroute. The Routeviews topologies are collected in every June and December from December 1997 until the end of May

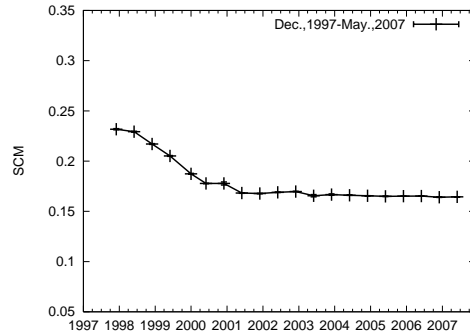


Fig. 1. SCM value of Routeviews topologies

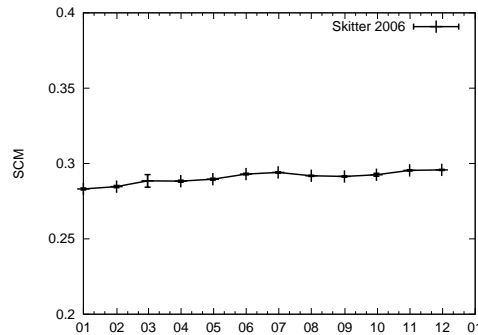


Fig. 2. SCM value of Skitter topologies

2007 (including snapshots in May 2007). Some instances are ignored due to the inconsistent scales such as those in December 1999. The Skitter graphs are more recent, collected in every month of 2006, with a scale around 9000 nodes.

#### A. Clustering Significance

First of all, we study the significance of clustering features of the AS-level topology. We use SCM as the metric and show the average value monthly in Fig. 1 and Fig. 2.

We observe that the SCM value for the Routeviews topologies has decreased steadily over the 9 years. From December 1997 to December 2000, the decreasing rate is much higher than the following years. After the topology size exceeds 10000 nodes (starting from the point of June, 2001), the decreased amount is negligible. Therefore, although the Internet is growing extremely fast, we expect the SCM value to be very stable in the future and no less than 0.15. As to the skitter topologies, the SCM values are very stable and much higher than those of Routeviews topologies.

#### B. Cluster Size Distribution

Secondly, we study the evolution of the cluster size distribution. Fig. 3 shows the Complementary Cumulative Distribution Function (CCDF) of the cluster size in log-log scales for the Routeviews topology instance on June

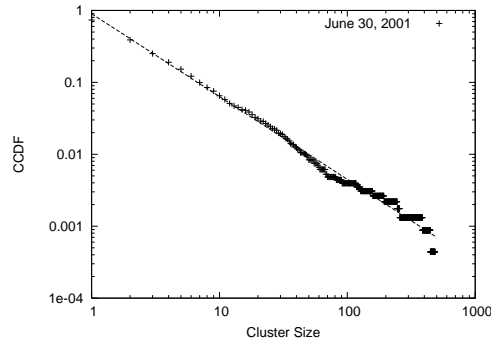


Fig. 3. CCDF of cluster size, Routeviews topology on June 30, 2001

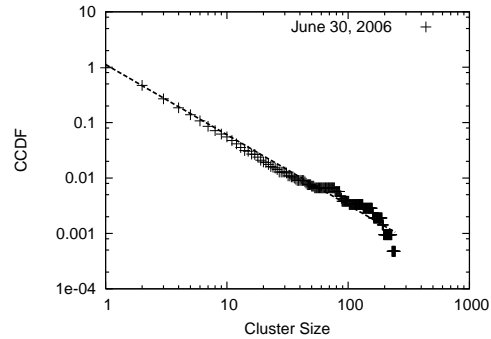


Fig. 4. CCDF of cluster size, Skitter topology on June 30, 2006

30, 2001. The cluster size follows a very skewed distribution: around 90% of clusters consist of less than 10 nodes while a few clusters have a huge cluster size of around 500 nodes. The high variability and skewed cluster size distribution also appear for skitter topologies, as shown in Fig. 4.

This type of distributions can be estimated by the well-known power law distributions in which two variables have linear correlation in log scales. For Fig. 3, we measure the correlation coefficient in log scales and observe a value around 0.98. The high linear correlation in log scales indicates our estimation on the cluster size distributions is valid. Using linear regression, the power law exponent is -1.22 and -1.3 for Routeviews and Skitter topologies respectively. All the other topology instances are measured in the same way. We find the correlation coefficients are no less than 0.97 for all the topology instances except for the Routeviews snapshots in December 1997 which have an average correlation coefficient of 0.957. Therefore, power law can be considered as a good estimation of the cluster size distribution. Next, we examine the evolution of the cluster size power law exponent.

Fig. 5 shows the cluster size power law exponent of Routerviews topologies in every June and December over the 9 years. As the topologies in Dec. 1999 are not consistent in scale, we use the snapshots in Jan. 2000 instead. We observe that the power law exponent only changes slightly over time. During the 9 years, the average exponent varies in a narrow range of -1.17 to -1.27, around a center of -1.20. For the skitter topologies, an even more stable exponent is observed in Fig. 6 because of the relatively short period.

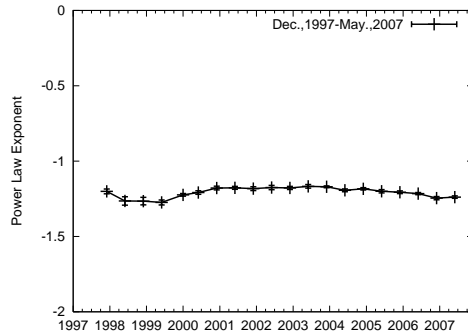


Fig. 5. Cluster size power law exponent of Routeviews topologies

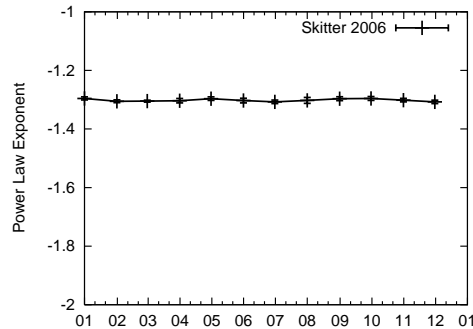


Fig. 6. Cluster size power law exponent of Skitter topologies

### C. Inter-cluster and Intra-cluster Connection

Now we take a close look at the topological structure of the AS-level Internet based on clusters, i.e., the inter-cluster and intra-cluster connection.

We first study the distribution of the inter-cluster connection which reflects the peering relationships between clusters. Fig. 7 shows the CCDF of the inter-cluster connection for the Routeviews topology on June 30, 2001. The high variability and the perfect linear correlation in log-log scales indicate a power law distribution. Through the measurement on both sets of topology instances, we claim the inter-cluster connections of Routeviews and Skitter snapshots follow power law distribution. Fig. 8 and Fig. 9 show the power law exponents for the two sets of topologies. As shown in Fig. 8, over the 9 year time period, the power law exponents of Routeviews topologies decrease slightly with the growth of topology size. For the year of 2006, the power law exponents of Routeviews topologies are lower than those of Skitter topologies.

The distributions of the intra-cluster connection follow power laws as well for all the tested topologies. Fig. 10 shows the CCDF of the intra-cluster connection of nodes in the Routeviews topology instance of June 30, 2001. Around 90% of the nodes have their intra-cluster connection less than 2 and a few nodes present high intra-cluster connection. As shown in the figure, the highest intra-cluster connection is no less than 500.

Now let us examine *the detailed topological structure of individual clusters*. We focus on the clusters with no less than 3 nodes, since the structures are deterministic for orphan nodes and 2-node clusters.

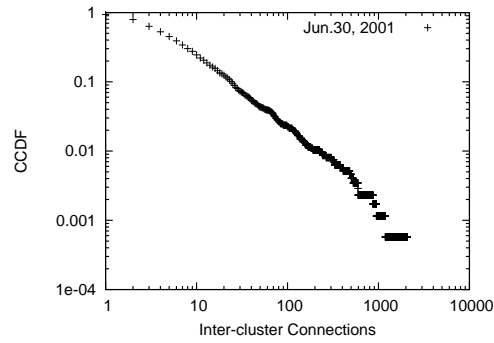


Fig. 7. CCDF of inter-cluster connection, Routeviews topology on June 30, 2001

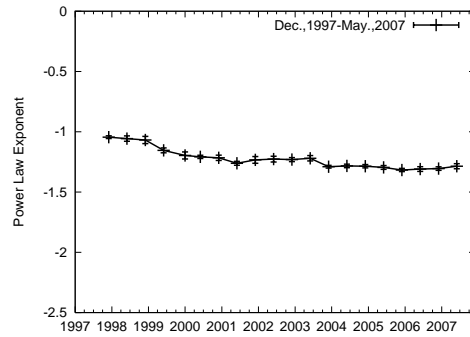


Fig. 8. Inter-cluster connection power law exponent of Routeviews topologies

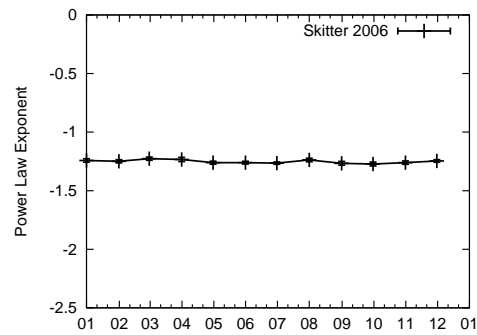


Fig. 9. Inter-cluster connection power law exponent of Skitter topologies

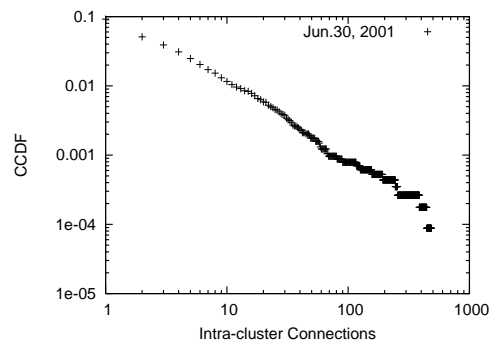


Fig. 10. CCDF of Intra-cluster connections, Routeviews topology on June 30, 2001.

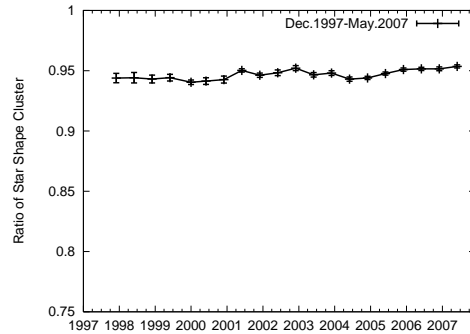


Fig. 11. Ratio of “stars” with no less than 3 nodes in Routeviews topologies.

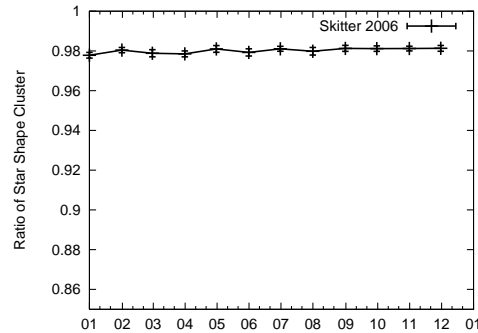


Fig. 12. Ratio of “stars” with no less than 3 nodes in Skitter topologies.

Among all the “non-trivial” clusters, the “star” structure is very common. In a “star” shaped cluster, one node appears as the star center and all the other nodes inside the cluster connect only with the center. Fig. 11 and Fig. 12 present the average ratio of “stars” with no less than 3 nodes for Routeviews and Skitter topologies. Around 95% of clusters are actually “stars” for Routeviews topologies and at least 98% for Skitter topologies. Moreover, we observe that the largest clusters in all the topologies have a “star” shape.

We further examine the rest of 5% clusters in Routeviews graphs which do not have a “star” shape. Our measurement leads to the following interesting discoveries:

- The “non-star” clusters are not dominating in size. The average size of the “non-star” clusters is limited in the range of  $[7, 11]$  and increases extremely slowly with time. During the 9 year period, the AS-level topology grows from 3000 nodes up to 25000 nodes, while the average size of the “non-star” clusters shows an increase of only 4 nodes.
- Most of the “non-star” clusters are loosely connected. We observe that only those “non-star” clusters with less than 5 nodes are close to a full mesh while in other clusters, the number of connections has linear correlation to the number of nodes, which is far from the densely connected full mesh.

#### D. Discussions

The above observations lead to an interesting and intuitive view about the AS-level Internet topology. Large substructures with very dense connections such as full mesh are not common in the AS-level topology. Instead, the AS-level topology can be simplified as a huge graph consists of a few loosely connected small meshes and thousands of inter-connected “stars” which have highly skewed and varied sizes.

Our observations are also consistent with the well established degree power law. As large clusters are all identified as “stars”, we can map the centers of those huge “stars” to the small number of nodes with extremely high degrees. Other nodes which are connected with the star centers correspond to the majority of nodes which have low degrees.

### V. EVALUATING TOPOLOGY MODELS

In this section, we compare the clustering features of the AS-level Internet topology with the generated graphs from different topology models. We show that under different construction approaches, the clustering features are very different. Our work in this section can be treated as an evaluation on different topology generators in capturing the clustering features of the Internet.

Along with the development of measurement techniques, knowledge about the Internet topology has become more and more comprehensive and accurate. Each stage in understanding the Internet topology is coupled with a set of topology models and generators which try to capture the newly discovered topological properties. In this section, four topology models are evaluated among which three (INET, BA, GLP) target on matching the degree power law of the Internet and one (Waxman) is designed for generating random topologies. We first review these four topology models.

#### A. Review of Topology Models

Waxman [26] is the first popular topology model used for network simulations. It takes Euclidean distance as the only parameter for connecting network nodes. Waxman is not considered as a realistic generation model for network simulations as it can neither capture the hierarchical structure of the Internet nor the power law degree distribution. We take Waxman in our study just for the purpose of providing a complete view of the clustering features in different topological structures.

The other three topology models are all degree-based, i.e., they try to match the degree power law of the Internet, but with different generation methods.

- The BA [4] model tries to emulate the growing process which leads to a degree power law. It starts with a small connected network core and sequentially adds new nodes and links to the existing network based on the *Linear Preferential Attachment* rule, i.e. new links are attached to an existing node with a probability proportional to the degree of existing nodes. As a consequence, high-degree nodes are always preferred to add new connections with, as leads to degree power law.

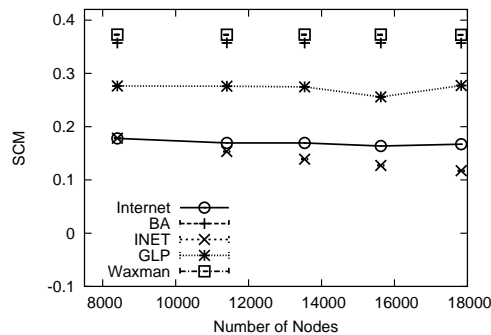


Fig. 13. SCM value of generated topologies

- The INET [13] generator also follows the linear preferential attachment but can generate more realistic degree exponents. It first calculates the degree power law distribution based on the measured degree exponent of the Internet. It then assigns degree to each node according to the given distribution. A spanning tree is constructed among all the nodes with degree more than 1 to guarantee a connected graph. At the end, additional links are added to match the given degree distribution using linear preferential attachment.
- The GLP [5] model uses a variant of the linear preferential attachment model to grow the network. Precisely, it gives higher preference to high degree nodes when adding new links.

In this set of experiments, we compare the clustering features of Routeviews topologies with generated graphs from the above four topology models. All the generated graphs have the same number of nodes as the Routeviews instances in every June from 2000 to 2004. For each model, at a given network size, 50 graph instances are generated to reduce the effects of randomness on topological properties.

### B. Clustering Significance

First, we examine the significance of clustering features by the SCM value of each generated topology.

Fig. 13 shows the SCM value of the generated graphs and the Routeviews instances. All the models produce quite stable SCM values. Compared with the Internet, graphs produced by Waxman, BA, and GLP have much higher SCM values. Only the INET model outputs topologies with SCM values close to the Internet. However, the deviation between the SCM of Internet and INET topologies increases with the topology size. The large difference in the SCM values of BA, GLP and INET also indicates different topological structures although all the three models generate degree power laws.

### C. Cluster Size Distribution

In this set of experiment, we compare the cluster size distribution of the generated topologies.

We find the cluster size of Waxman, BA and GLP graphs do not exhibit high variability and skewness as the Internet snapshots, while the cluster size of INET topologies follow power law distribution. Thus we compare the power law exponents of INET with the Internet.

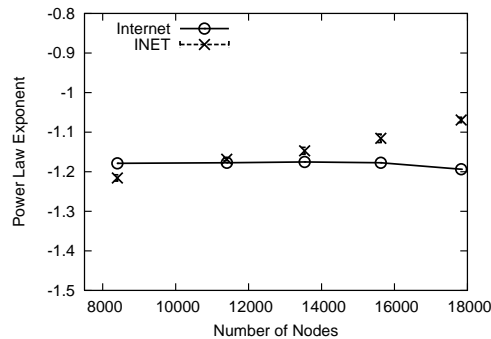


Fig. 14. Cluster size power law exponent of INET

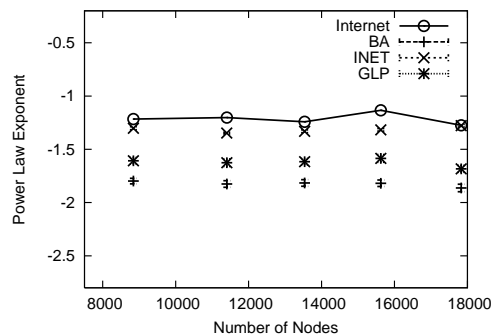


Fig. 15. Inter-cluster connection power law exponent of generated topologies

Fig. 14 shows the comparison of cluster size exponents between INET and the Internet. As shown in the figure, the exponents of INET are slightly different from those of the Internet. With the increase of topology size, the power law exponent of INET increases significantly and the distance to the exponents of the Internet is enlarged. However, comparatively speaking, this model behaves much better in capturing the cluster size distribution of the Internet than the other topology models.

#### D. Inter-cluster/Intra-cluster Connections

We also examine the four topology models in terms of inter-cluster/intra-cluster connections. We first compare the inter-cluster connection distribution of each model with the Internet snapshots.

We find except for Waxman, all the models can generate power law inter-cluster connection distributions. We then compare the power law exponents of these models in Fig. 15. Although slightly lower than the exponents of the Internet, the power law exponents of INET are more accurate than those of BA and GLP.

Before we evaluate the intra-cluster connection distributions, we examine the “star” structure in the generated graphs. We find the “star” clusters are very common in all the generated topologies. The ratio of “stars” for each model is even higher than in the Internet topologies. The high ratios of “stars” cause the distribution of intra-cluster connection consistent with the cluster-size distribution. Therefore, only INET can generate graphs which have power law intra-cluster connection distributions. We compare the exponents of INET with those of the Internet as shown

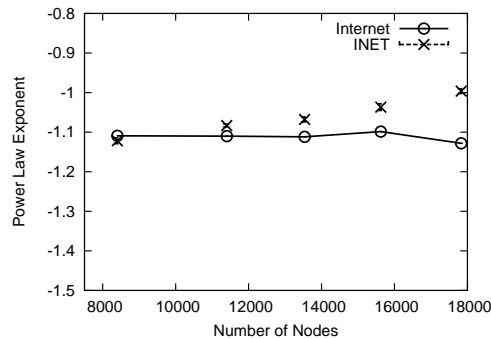


Fig. 16. Intra-cluster connection power law exponent of INET

in Fig. 16. We find that similar to cluster size distribution, the power law exponents of the two sets of topologies are close but the gap between them increases significantly with the topology size.

Here we give a quick summary of our findings: different topology models produce different clustering substructures. Among all the investigated models, the INET generator shows the best performance in matching the clustering features of the Internet. On the other hand, however, there is still a certain gap in the “matching”. Thus, in the next section, we present our new topology model, and demonstrate its power to reproduce clustering features.

## VI. CBTM: CLUSTERING BASED TOPOLOGY MODEL

Clustering features incorporate both local topological properties and global characteristics. Therefore, we expect that by reproducing the clustering features of the Internet instances, we can capture the commonly used topological properties of the reference graphs. Following this logic, we develop a new topology generation model which can match not only the clustering features of the Internet topology but also the other well known properties such as the degree power law and the small world behavior, naturally due to its capacity of generating realistic clustering features.

### A. Model

We name the new model as **Clustering Based Topology Model** (CBTM). As indicated by its name, CBTM is based on the statistics of clustering features from the Internet topology instances. The main metrics used in this model include cluster size distribution and inter/intra-cluster connection distribution. The generation process is illustrated by Fig. 17, which can be broken down into the following four steps.

- 1) Create clusters with various number of nodes based on the input number of nodes and the cluster size distribution, as shown in Fig. 17(a).
- 2) Intra-connect nodes within each cluster as a “star”. According to our study, “star” shaped clusters are dominating in all the Internet instances. To simplify the process, we ignore the small portion of “non-stars”. The resultant graph is shown in Fig. 17(b).

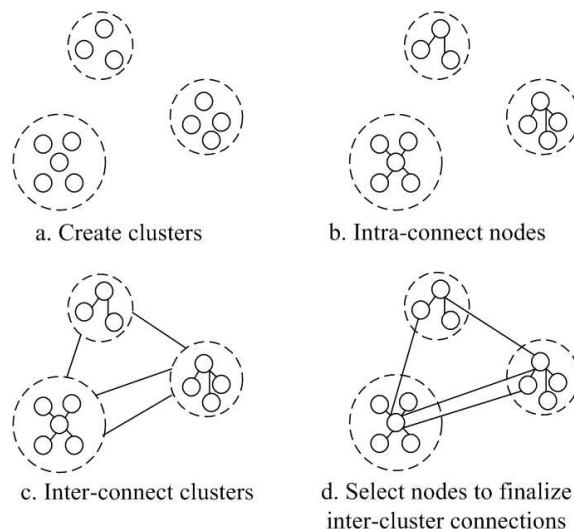


Fig. 17. Topology generation process in CBTM

- 3) Inter-connect clusters. As in Fig. 17(c), after clusters are created internally, add inter-cluster connections based on the distribution of inter-cluster connection and the correlation between pair of clusters. For a particular size, the number of inter-cluster connections can be approximated as a truncated power law distribution.
- 4) Select nodes within each cluster to finalize the inter-cluster connections, as shown in Fig. 17(d).

One of the key steps in this process is how to determine a pair of clusters for an inter-cluster connection. We need to find out the correlations between neighbor clusters. Intuitively, the pairwise correlation between clusters are related to the cluster size. A big cluster is more likely to have a large neighbor since big clusters have high inter-cluster connections. On the other hand, small clusters are dominating in numbers, so there is also a non-trivial probability to select a small cluster as a neighbor. We therefore take cluster size as the major parameter to characterize the pairwise correlation between clusters.

Another question is how to pick a node for an inter-cluster connection. Obviously, the inter-cluster connections are not evenly shared among the nodes within a cluster. The allocation of inter-cluster connections also affects the power law distribution of node degree. Based on our observation, the probability of attaching an inter-cluster connection with the center of a “star” cluster is proportional to the size of that cluster. Therefore, large clusters which have high inter-cluster connections connect with other clusters by its center in most cases, as also helps to lead high skewed degree distribution.

## B. Evaluation

1) *Clustering Metrics:* Now we compare the clustering statistics of the topologies generated by CBTM with Routeviews Internet snapshots and graphs generated from INET.

We first measure the clustering features including SCM value, cluster size distribution, and inter/intra-cluster connection distributions. Fig. 18 plots SCM value at different topology scales. From this figure, we can see that

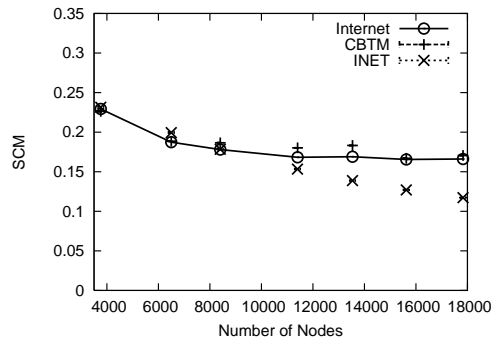


Fig. 18. SCM value of CBTM

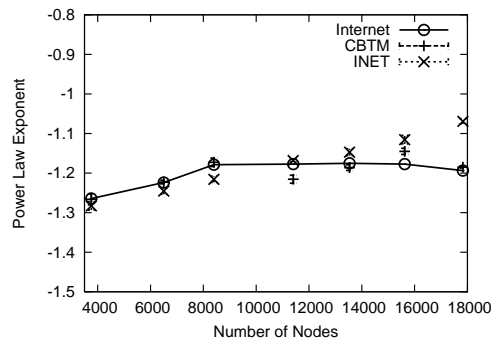


Fig. 19. Cluster size power law exponent of CBTM

CBTM shows very similar clustering significance to the reference topologies. The evolution trend of CBTM is also stable and comparable to the Internet instances. Compared with CBTM, INET departs from the reference topologies with steadily decreasing clustering significance.

Then we examine the cluster size distribution of the generated graphs. All CBTM topologies show highly skewed distributions which follow power laws. Fig. 19 illustrates the power law exponents of the three sets of topologies. Again, CBTM beats INET by its more accurate size exponents. In this set of experiment, INET shows higher exponents than CBTM and Internet when topology size gets large. By closely examining the cluster size distributions, we find that the top clusters in terms of cluster size in INET are much smaller than the top clusters in CBTM and Routeviews snapshots, while more middle sized clusters with 10 to 100 nodes are generated in INET.

Next, we evaluate CBTM in terms of the Inter-/Intra-cluster connection distributions. The results are plotted in Fig. 20 and Fig. 21 respectively. For both metrics, the three sets of topologies all keep power law distributions but with different exponents, and it is clear that CBTM has better performance than INET.

2) *Other Topological Metrics:* Besides the clustering related metrics, we also measure some most commonly used graph properties.

**Degree Distribution** First, we show the degree distributions for the three topology sets. As the three sets of topologies all have power law degree distributions, we only compare the power law exponents.

Fig. 22 shows the degree power law exponents. Except for the last reference topology, INET topologies always

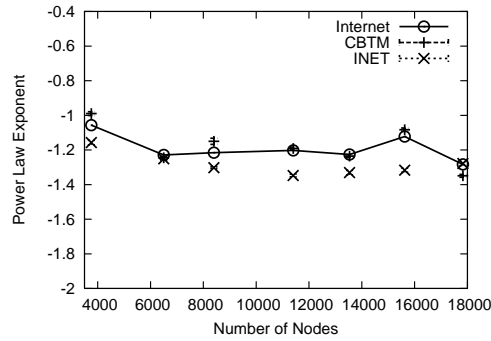


Fig. 20. Inter-cluster connection power law exponent of CBTM

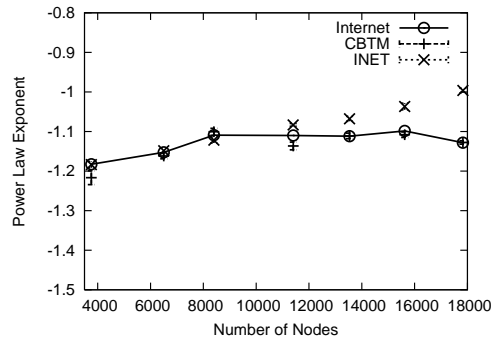


Fig. 21. Intra-cluster connection power law exponent of CBTM

have lower power law exponents than Internet and CBTM. These results indicate that clustering metrics can effectively capture degree distributions.

**Small World Behavior** Two metrics are used to evaluate small world behaviors: Characteristic Path Length and Clustering Coefficient. Now we compare CBTM with Internet and INET against these two metrics.

Fig. 23 shows the characteristic path length of the Internet instances and the generated graphs. Topologies from both CBTM and INET show lower values than Routeviews snapshots. However, CBTM can still capture this metric better than INET and shows much better performance compared with the other models.

The clustering coefficients of the three sets of topologies are shown in Fig. 24. The coefficients of CBTM are

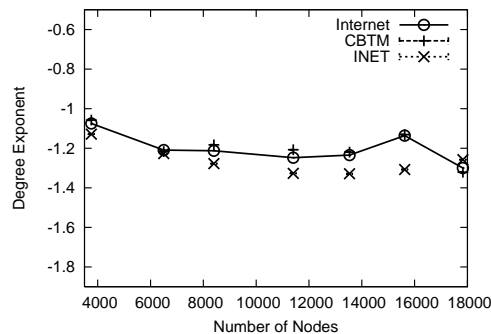


Fig. 22. Node degree power law exponent of CBTM

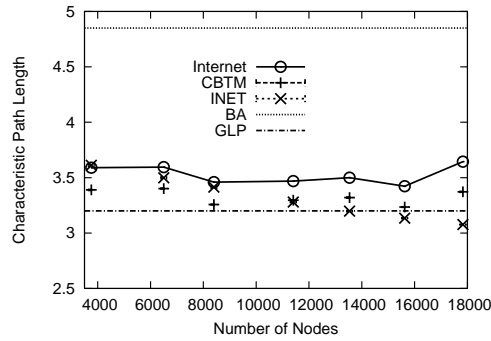


Fig. 23. Characteristic path length of CBTM

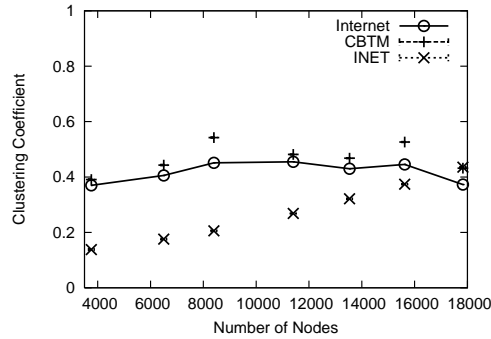


Fig. 24. Clustering coefficient of CBTM

slightly higher than the reference topologies, but are much more accurate than those from INET. Combining the observations in the characteristic path length, the structures of the INET graphs are more random than the Internet AS-level topologies while CBTM graphs show less randomness than INET, which is more similar to the reference topologies.

**Likelihood** The last topology metric we measure is assortativity coefficient [22], which is shown in Fig. 25. Assortativity coefficient quantifies how likely nodes with similar degrees are connected. In networks with negative assortativity coefficients, nodes tend to connect with others with different degrees. Our results show that CBTM can produce accurate assortativity coefficients.

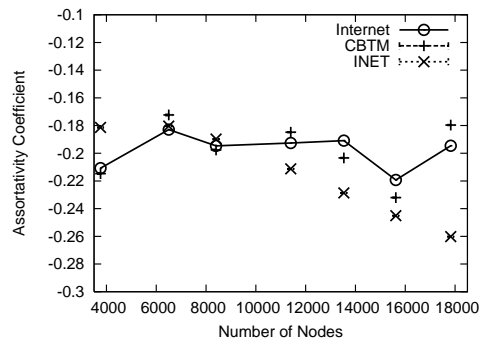


Fig. 25. Assortativity coefficient of CBTM

3) *Discussions*: The above measurement studies evaluate CBTM in a comprehensive way, from newly proposed clustering features to well accepted topology attributes. From the results, we can claim CBTM is a more realistic topology model than others. We also demonstrate that clustering features are important topological attributes which are highly related to both global and local properties of a network topology. They can well reflect global hierarchical structure of the Internet topology as well as local properties such as node degrees and joint degree correlations.

## VII. RELATED WORK

The discovery of degree power law in [10] has inspired a big wave of research efforts in understanding the Internet topology and its evolution [5], [8], [12], [18]–[20]. A number of topology models have been proposed to reproduce the observed degree distributions (as mentioned in Section V). However, degree distribution is too limited to capture the real structure of the Internet topologies, especially the global properties: graphs with the same degree distribution may vary significantly. In [5], Bu et. al. find out the AS-level Internet topologies have “small world” behaviors [25] which is described by the metrics of clustering coefficient and characteristic path length. Using these two metrics, some representative topology models are evaluated and it turns out that none of them can capture the small world behavior of the Internet. Most of those topology models output low clustering coefficient. This is the first attempt in clustering features exploration in the Internet topology and hence there exist some limitations as discussed in the introduction section.

In more recent work [18], degree correlations are taken as main metrics to study the Internet topologies. The authors use a series of degree correlation distributions (*dK-series*) within subgraphs of increasing sizes to characterize Internet topologies. It is proved that *dK-series* can capture all the graph properties including those proposed in the future when  $d$  is big enough. In the future work, it would be interesting to investigate how big is  $d$  in order to capture all the clustering properties we have characterized.

Different from the aforementioned research work, in [15], graph property statistics are not playing as key roles in topology modelling and evaluation. This work considers practical technological constrains and economic considerations as the main forces of the router-level topology design. We are interested in extending our clustering feature characterization into router-level topologies, and examine the applicability of CBTM at router level.

## VIII. CONCLUSIONS

In this paper, we have explored the clustering features of the AS-level Internet topology through extensive measurements. To obtain a quantitative view, we propose a set of new metrics which cover all levels of clustering features. We also evaluate the clustering features presented in the generated graphs from different topology models. We observe the existing models such as BA, INET and GLP, even though can match the degree power law of the Internet, behave very differently in producing the clustering features. This observation shows the strength of the clustering features in distinguishing different topological structures and models.

Through the clustering feature characterization, we can see the AS-level Internet topology is well structured and can be described effectively by the clustering features. As highly correlated with other topological properties such as node degree distribution and hierarchical structure, clustering features provide big potential for realistic topology modelling. Thus we design a simple while accurate topology model, CBTM, based on the clustering features characterization. Through extensive evaluations, we find that CBTM can reproduce the clustering features and other well accepted topology attributes for the AS-level Internet topology.

## REFERENCES

- [1] Cooperative association for internet data analysis. <http://www.caida.org>.
- [2] Oregon routeviews project. <http://www.routeviews.org/>.
- [3] Public route server and looking glass site list. <http://www.traceroute.org>.
- [4] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.
- [5] T. Bu and D. Towsley. On distinguishing between internet power law topology generators. In *Proceedings of IEEE INFOCOM*, pages 638–647, June 2002.
- [6] J.-H. Cui, M. Faloutsos, D. Maggiorini, M. Gerla, and K. Boussetta. Measuring and modelling the group membership in the internet. In *Proceedings of ACM SIGCOMM/USENIX Internet Measurement Conference (IMC)*, pages 65–67, October 27-29, 2003.
- [7] J. S. Deogun, D. Kratsch, and G. Steiner. An approximation algorithm for clustering graphs with dominating diametral path. *Information Processing Letter*, 61(3):121–127, 1997.
- [8] X. Dimitropoulos, D. Krioukov, M. Fomenkov, and B. Huffaker. As relationships: Inference and validation. *ACM SIGCOMM Computer Communication Review*, pages 29–40, 2007.
- [9] D. Estrin, Y. Rekhter, and S. Hotz. Scalable inter-domain routing architecture. In *Proceedings of ACM SIGCOMM*, pages 40–52, 1992.
- [10] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of ACM SIGCOMM*, pages 251–262, August 1999.
- [11] P. Fraigniaud. A new perspective on the small-world phenomenon: Greedy routing in tree-decomposed graphs. In *Proceedings of 13th Annual European Symposium on Algorithms (ESA)*, pages 791–802, 2005.
- [12] C. Gkantsidis, M. Mihail, and E. Zegura. Spectral analysis of internet topologies. In *Proceedings of IEEE INFOCOM*, March 2003.
- [13] C. Jin, Q. Chen, and S. Jamin. Inet: Internet topology generator. *Technique Report CSE-TR-433-00, University of Michigan, EECS Department*, 2000.
- [14] D. Krioukov, K. Fall, and X. Yang. Compact routing on internet-like graphs. In *Proceedings of IEEE INFOCOM*, 1:–219, March 2004.
- [15] L. Li, D. Alderson, W. Willinger, and J. C. Doyle. A first-principles approach to understanding the internet’s router-level topology. In *Proceedings of ACM SIGCOMM*, pages 3–14, 2004.
- [16] Y. Li, L. Lao, and J.-H. Cui. SACA: SCM-based Adaptive Clustering Algorithm. In *Proceedings of IEEE MASCOTS*, pages 271–279, September 2005.
- [17] Y. Li, L. Lao, and J.-H. Cui. SDC: A Distributed Clustering Protocol for Peer-to-Peer Networks. In *Proceedings of IFIP Networking*, May 15 - 19 2006.
- [18] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat. Systematic topology analysis and generation using degree correlations. In *Proceedings of ACM SIGCOMM*, pages 135–146, 2006.
- [19] P. Mahadevan, D. Krioukov, M. Fomenkov, and B. Huffaker. The internet as-level topology: Three data sources and one definitive metric. *ACM SIGCOMM Computer Communication Review*, 36(1):17–26, 2006.

- [20] A. Medina, A. Lakhina, I. Matta, , and J. Byers. Brite: Universal topology generation from a user’s perspective. In *Proceedings of International Workshop on Modeling, Analysis and Simulation of Computer and Telecommunications Systems (MASCOTS)*, October 2001.
- [21] A. Medina, I. Matta, and J. Byers. On the origin of power laws in internet topologies. In *SIGCOMM Comput. Commun. Rev.*, volume 30, pages 18–28, New York, NY, USA, 2000. ACM Press.
- [22] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, October 2002.
- [23] L. Ramaswamy, A. Iyengar, L. Liu, and F. Douglis. Connectivity based node clustering in decentralized peer-to-peer networks. In *3rd International Conference on Peer-to-Peer Computing*, 2003.
- [24] S. van Dongen. A cluster algorithm for graphs. *Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam*, May 2000.
- [25] D. J. Watts and S. H. Strogatz. Collective dynamics of “small-world” networks. *Nature*, 393:440–442, June 1998.
- [26] B. M. Waxman. Routing of multipoint connections. *IEEE Journal on Selected Areas in Commucations*, 6:1617–1622, December 1988.