Abstract—Modeling peer-to-peer (P2P) networks is a challenge for P2P researchers. It is the key to provide insight into the nature of the underlying system and building a successful model able to generate realistic topologies for simulation purpose. In this paper, we provide a detailed analysis of large-scale P2P network topology, using Gnutella as a case study. First, we re-examine the power-law distributions of the Gnutella network discovered by previous researchers. Our results show that the current Gnutella network deviates from the early power-laws, suggesting that the Gnutella network topology may have evolved a lot over time. Second, we identify important trends with regard to the evolution of the Gnutella network between September 2005 and February 2006. Third, we provide a novel two-layered approach to study the topology of the Gnutella network. Due to the limitations of the power-laws, we divide the Gnutella network into two layers, namely the mesh and the forest, to model the hybrid and highly dynamic architecture of the current Gnutella network. We give a detailed analysis of the topology of the mesh and present four power-laws concerning the mesh topology. Moreover, we examine the topology properties of the forest and provide one empirical law concerning the tree size. Using the two-layered approach and laws proposed, we can generate realistic topologies easily.

I. INTRODUCTION

Modeling the topologies of peer-to-peer (P2P) networks is an important open problem. An accurate topological model can have significant influence on P2P research. First, we can gain detailed insight into the nature of the underlying system. Second, the model can enable analytical analysis of algorithms and facilitate design of more efficient protocols that take advantage of topology properties. Third, we can generate more accurate artificial topologies for simulation purpose. Further more, we can predict future trends and thereby address potential problems in advance.

In this paper, we provide a detailed analysis of large-scale P2P network topology, giving results concerning major topology properties and main distributions. In our study, we choose Gnutella as a case study, as it has a large user community and open architecture. Our work can be summarized by the following points.

First, we re-examine the power-law distributions of the Gnutella network discovered by previous researchers. Our results show that the current Gnutella network deviates from the early power-laws. This observation suggests that the Gnutella network topology may have evolved a lot over time.

Second, we identify important trends with regard to the evolution of the Gnutella network between September 2005 and February 2006.

As our primary contribution, we provide a novel two-layered approach to study the topology of the Gnutella network. Due to the limitations of the power-laws, we divide the Gnutella network into two layers, namely the mesh and the forest, to model the hybrid and highly dynamic architecture of the current Gnutella network. We give a detailed analysis of the topological properties of the mesh and present four power-laws concerning the mesh topology. Moreover, we examine the topology properties of the forest and provide one empirical law concerning the tree size.

Finally, we focus on the generation of realistic topologies using our approach and laws proposed.

The rest of this paper is organized as follows. Section II presents background and previous work. In Section III, we present our instances of the Gnutella network. In Section IV, we re-examine the power-law distributions discovered by previous researchers, identify the trends concerning the evolution of Gnutella network. In Section V, we analyze the limitations of the power-laws and introduce our new two-layered approach to study the topology of Gnutella network. In Section VI, we analyze the topological properties of the mesh and present four power-laws concerning the mesh topology. In Section VII, we examine the topology properties of the forest and provide one empirical law concerning the tree size. In Section VIII, we discuss the practical uses of our approach and laws. Finally, Section IX concludes our work.

II. BACKGROUND AND PREVIOUS WORK

A. Gnutella Protocol and the Crawler

Gnutella protocol 0.4 [1] employs a pure decentralized model. In this model, individual nodes, also called servants, are equal in terms of functionality. They not only perform server-side roles such as matching incoming queries against their local resources and respond with applicable results, but also offer client-side functions such as issuing queries and collecting search results. All servants are connected to each other randomly. Figure 1 illustrates the topology of the Gnutella 0.4 network.

Gnutella protocol 0.6 [2] employs a hybrid architecture combining centralized and decentralized model. Servents are categorized into leaf and ultrapeer. A leaf keeps only a small
number of connections to ultapeers. An ultapeer maintains connections with other ulpaqueers and acts as a proxy to the Gnutella network for the leaves connected to it. An ultapeer only forwards a query to a leaf if it believes the leaf can answer it, and leaves never relay queries between ultapeers. Figure 2 illustrates the topology of the Gnutella 0.6 network. Protocol 0.6 is compatible with protocol 0.4, which implies that the current Gnutella network can contain some fraction of nodes of former protocol specification 0.4.

We developed a crawler to collect topology information of the Gnutella network, based on message communication mechanism of both protocol 0.4 [1] and protocol 0.6 [2]. The crawler is based on the Limewire [3] open source client and performs a breadth first searching on the network in parallel. It can discover more than 100,000 nodes in half an hour.

B. Power-law

Power-laws have been found in numerous diverse fields spanning sociological, geological, natural and biological systems. Power-laws of the form \( y \propto x^\alpha \) enables a compact characterization of topologies through their exponents. Faloutsos et al. [4] discovered four power-laws characterizing the topology of the Internet, while Magoni et al. [5] found another four power-laws of the Internet.

In [6] [7] and [8], several power-laws were found with regard to the topology of the Gnutella network. P2P studies usually assume that these power-laws characterize the topology of P2P networks and use synthetically generated topologies following these power-laws, although Ripeanu et al. [9] argued that the connection distribution of the more recent Gnutella network deviates from a pure power-law. In Section IV, we will re-examine these power-laws.

III. OUR GNUTELLA NETWORK INSTANCES

We have studied the topology of the Gnutella network from September 2005 until February 2006. We can build the graph of nodes by analyzing the collected data on the Gnutella network. We model two adjacent nodes that have at least one connection between each other by an edge. We treat the Gnutella network as a undirected graph.

In this paper, we provide two snapshots of the Gnutella network that correspond to five-month intervals approximately, namely the 091505 instance and the 021106 instance. In Table I, we present some basic statistics about our instances and previous work [6] [7]. In Table I, \( l \) represents the average shortest distance and \( k \) represents the average degree.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>091505</td>
<td>021106</td>
<td>V34206</td>
</tr>
<tr>
<td>Time</td>
<td>09/05</td>
<td>02/06</td>
<td>09/03</td>
</tr>
<tr>
<td>Nodes</td>
<td>118,187</td>
<td>112,925</td>
<td>34,206</td>
</tr>
<tr>
<td>Edges</td>
<td>130,612</td>
<td>118,925</td>
<td>80,276</td>
</tr>
<tr>
<td>( l )</td>
<td>6.4</td>
<td>7.9</td>
<td>5.4</td>
</tr>
<tr>
<td>Diam.</td>
<td>22</td>
<td>24</td>
<td>16</td>
</tr>
<tr>
<td>( k )</td>
<td>2.20</td>
<td>2.20</td>
<td>2.57</td>
</tr>
</tbody>
</table>

Fig. 1. Topology of the Gnutella 0.4 Network.

Fig. 2. Topology of the Gnutella 0.6 Network.

IV. CURRENT GNUTELLA NETWORK TOPOLOGY

In this section, we examine the power-laws of the Gnutella network described in previous literatures against our two instances. Note that in this paper we only present the examination of two early power-laws that give strong evidence for the representativeness of a topology [10]. The goal of our work is to find out whether the topology of the current Gnutella network accords with the early power-laws.

We use linear regression to fit a line in a set of two-dimensional points using the least-square errors method. The validity of the approximation is quantified by the correlation coefficient ranging from -1.0 and 1.0. The absolute value of the correlation coefficient is ACC. An ACC value of 1.0 indicates perfect linear correlation. In general, the ACC level should be great than 0.90 to validate linear correlation.

A. Rank Distribution

In this section, we study the degrees of the nodes in the Gnutella network.

Power-Law of Rank Exponent \( \mathcal{R} \): The degree \( d_v \) of a node \( v \) is proportional to the rank of the node \( r_v \) to the power of a constant \( \mathcal{R} : d_v \propto r_v^\mathcal{R} \). The rank \( r_v \) of a node \( v \) is defined as its index in the order of decreasing degree.

Jovanovic [6] found that the early Gnutella network followed the above power-law with rank exponent of -0.98 and ACC of 0.94. For our two instances, the rank exponent is -0.64268 and -0.60681 and ACC is 0.92178 and 0.88120 in chronological order as we see in Figure 3. The low ACC values imply that this power-law is relatively weak in the 091505 graph and even invalided in the 021106 graph.

Compared with a pure power-law distribution, the two graphs deviate from the linear regression with similar patterns. On the one hand, the nodes with high rank are of too small degree. This is because the Gnutella protocol 0.6 imposes a limit on maximal connections of an ultapeer. On the other hand, there are too many nodes with degree around 30,
resulting the curve breakouts from the linear regression. This pattern suggests that ultrapeers in the Gnutella 0.6 network tend to have the connection limit around 30. Moreover, the 021106 graph is somewhat different from the 091505 graph. First, the nodes with high rank in the former graph are of smaller degree compared with the counterparts in the latter, implying that protocol 0.6 is effectively replacing protocol 0.4. Secondly, the curve after a degree of approximately 30 drops much more sharply in the former graph than in the latter, which suggests that ultrapeers tend to employ as many connections as they can.

B. Degree Distribution

In this section, we study the distribution of the degrees of the nodes. Note that the degree power law we present here is different from the one in earlier work [6]. However, they both refer to the same distribution. The difference is that the former uses the cumulative probability distribution function, while the latter uses the probability distribution function. As a result, the exponents of the two power-laws differ approximately by one. The cumulative distribution is preferable because it can be estimated in a statistically robust way.

**Power-Law of Degree Exponent** $D$: The CCDF $D_d$ of a degree $d$, is proportional to the degree to the power of a constant $D : D_d \propto d^D$. The complementary cumulative distribution function (CCDF) of a degree $d$ is the percentage of nodes that have degree greater than the degree $d$.

Jovanovic [6] showed degree exponent of -1.4 and ACC of 0.96 for the early Gnutella network by probability distribution. Chen et al. [7] argued that cumulative probability distribution of the node degree follows a power-law. But they did not provide any information about the ACC value and the exponent value. For our two instances, the degree exponent is -2.25926 and -2.31074 and ACC is 0.91744 and 0.87718 in chronological order as we see in Figure 4. Again, the low ACC values imply that this power-law is relatively weak in the 091505 graph and even invalidate in the 021106 graph.

![Fig. 4. Log-log plot of $D_d$ versus the degree $d$.](image)

Compared with a pure power-law distribution, the graphs share some common patterns. There are too many nodes with degree around 30, and the resulting curves deviate from the linear regression. This is coincident with what we found in rank distribution.

Furthermore, in the 021106 graph, degrees in interval 5 to 20 follow an almost constant distribution, which means there are too few ultrapeers with a degree in this interval. This confirms our previous conclusion that ultrapeers try to hold more connections up to the limit. The curve of higher degree in the 021106 graph drops much more sharply, which agrees with our previous comment that the Gnutella protocol 0.6 prevents ultrapeers from employing a large number of connections.

V. THE TWO-LAYERED APPROACH

In this section, we first discuss the limitations of the power-laws and then present a new approach to study the topology of the Gnutella network.

A. Limitations of the Power-laws

Previous researches [11] and [10] suggest two key causes for power-law distributions in network topologies: incremental growth and preferential connectivity. **Incremental growth** refers to open networks that form by the continual addition of new nodes, and thus the gradual increase in the size of the network. **Preferential connectivity** refers to the tendency of a new node to connect to existing nodes that are highly connected or popular.

The topology of the Gnutella network is highly dynamic, since a node can join or leave the Gnutella network at any time. More specifically, most leaves tend to disconnect from the Gnutella network in several minutes after they connect to the network. The transient life-time of the leaves works against incremental growth. Moreover, due to the hybrid architecture of Gnutella protocol 0.6 [2], a leaf keeps only a small number of connections to ultrapeers and cannot connect to other leaves. This limitation on leaves also works against preferential connectivity, because leaves can never become highly connected. Combining the above factors, we can explain why the current Gnutella network does not follow the early power-law distributions. It is the limitations of the power-laws that make them inappropriate for modeling hybrid and highly dynamic topologies.

As we mentioned earlier, P2P studies usually use synthetically generated topologies characterize by the early power-laws. These topologies may not reflect properties of current P2P networks. So there should be a new approach to model current P2P networks.

B. Our Approach

In our study, we propose a new two-layered approach to model the topology of the current Gnutella network. We split the Gnutella network into two layers, namely the mesh and the forest.

Before we present the analysis of our approach, we provide below a few definitions. Note that Magoni et al. [5] proposed some definitions to describe the AS network. We keep these definitions and modify them into the following ones. Figure 5 shows different kinds of nodes in a sample graph.

- Cycle node: a node that belongs to a cycle (i.e. it is on a closed path of disjoint nodes; in Figure 5, there are eleven cycle nodes).
A. Mesh Node Rank Exponent $\mathcal{R}_m$

In this section, we study the degrees of the nodes in the mesh. We sort the nodes in the mesh in decreasing order of degree $d_{v_m}$ and define the mesh node rank $r_{v_m}$ as the index of the node in the sequence. We plot the $(d_{v_m}, r_{v_m})$ pairs in log-log scale. The plots are shown in Figure 6. The data values are represented by points, while the solid lines represent the least-squares approximation.

![Fig. 6. Log-log plot of the mesh node degree $d_{v_m}$ versus the rank $r_{v_m}$ in the sequence of decreasing degree.](image)

The points of Figure 6 are well approximated by the linear regression. The ACC is 0.96425 for the 091505 instance and 0.96580 for the 021106 instance. This leads us to the following power law and definition.

**Power-Law 1 (Mesh Node Rank Exponent):** The degree $d_{v_m}$ of a mesh node $v_m$ is proportional to the rank of the mesh node $r_{v_m}$ to the power of a constant $\mathcal{R}_m$:

$$d_{v_m} \propto r_{v_m}^{\mathcal{R}_m}.$$  

**Definition 1:** Let us sort the mesh nodes of a graph in decreasing order of degree. We define the mesh rank exponent $\mathcal{R}_m$ to be the slope of the plot of the degrees of the mesh nodes versus the rank of the nodes in log-log scale.

**B. Mesh Node Degree Exponent $O_m$**

In this section, we study the distribution of the degrees of the nodes in the mesh. We define the frequency $f_{d_m}$ of a mesh node degree $d_m$ as the number of nodes in the mesh with degree $d_m$. We plot the $(f_{d_m}, d_m)$ pairs in log-log scale in Figure 7. In these plots, we exclude a small percentage of nodes of higher degree that have frequency of one, but still plot 99.9% of the total number of nodes. As we saw earlier, the higher degrees are described and captured by the mesh.

![TABLE II

<table>
<thead>
<tr>
<th>Stat.</th>
<th>091505</th>
<th>021106</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>16,487</td>
<td>11,852</td>
</tr>
<tr>
<td>p(m)</td>
<td>15.4%</td>
<td>10.0%</td>
</tr>
<tr>
<td>Edges</td>
<td>27,467</td>
<td>23,539</td>
</tr>
<tr>
<td>l</td>
<td>5.2</td>
<td>6.5</td>
</tr>
<tr>
<td>k</td>
<td>3.33</td>
<td>3.97</td>
</tr>
<tr>
<td>Diam.</td>
<td>14</td>
<td>17</td>
</tr>
</tbody>
</table>

**VI. MESH TOPOLOGY ANALYSIS**

In this section, we study the topology properties concerning the mesh in the Gnutella network. In Table II, we present some basic statistics about the mesh in our instances. In Table II, $p(m)$ represents the percentage of nodes in the mesh, $l$ represents average shortest distance, and $k$ represents average degree.

- **Bridge node:** a node which is not a cycle node and is on a path connecting 2 cycle nodes (in Figure 5, there is one bridge node).
- **In-mesh node:** a node which is a cycle node or a bridge node (in Figure 5, the mesh has twelve in-mesh nodes).
- **In-tree node:** a node which is not an in-mesh node (i.e. it belongs to a tree; in Figure 5, each tree has four in-tree nodes).

Mesh is the set of in-mesh nodes and forest is the set of in-tree nodes.

- **Branch node:** an in-tree node of degree at least 2.
- **Leaf node:** an in-tree AS of degree 1.
- **Root node:** an in-mesh node which is the root of a tree.
- **Relay node:** a node having exactly 2 connections.
- **Border node:** a node located on the diameter of the network.

If we split the Gnutella network into the mesh and the forest, we can analyze the topological properties of the mesh and the forest respectively.

After a careful comparison between Figure 2 and Figure 5, we can find that the mesh in Figure 5 is composed merely of ultrapeers and acts as the backbone of the Gnutella network. Since ultrapeers are relatively stable and tend to stay in the Gnutella network for a longer time, it can meet the requirement of incremental growth. Further more, since ultrapeers can connect to other ultrapeers, it can meet the requirement of preferential connectivity. Hence, the topology of the mesh theoretically should comply with power-laws (see Section VI for detailed validation). On the other hand, we can also obtain major topology properties and distributions of the forest (see Section VII).

With the knowledge of both the topology of the mesh and the topology of the forest, we can model the topology of the Gnutella network easily by merging these two layers.

![Fig. 5. Different kinds of nodes.](image)
**Power-Law 2 (Mesh Node Degree Exponent):** The frequency $f_{dn}$ of a mesh node degree $d_m$, is proportional to the degree to the power of a constant $C_m$:

$$f_{dn} \propto d_m^{C_m}.$$  

**Definition 2:** We define the mesh node degree exponent $C_m$ to be the slope of the plot of the frequency of the mesh node degrees versus the degrees in log-log scale.

**C. Mesh Pair Rank Exponent $P_m$**

In this section, we study the **Number of distinct Shortest Paths (NSP)** of each pair of vertices in the mesh. The number of distinct shortest paths between two vertices is the number of shortest paths such as any of these paths have at least one vertex not in common [5]. The distribution of NSP is useful for evaluating the amount of redundancy edges involved in shortest path. Higher NSP values mean that if one edge of a shortest path between a pair of nodes is removed, there is still a probability for another shortest path of the same length to exist for this pair. We sort the pairs in the mesh in decreasing NSP $n_{pm}$ and define the pair rank $r_{pm}$ as the index of the pair in the sequence. We plot the $(n_{pm}, r_{pm})$ pairs in log-log scale. The plots are shown in Figure 8. Due to the enormous amount of node pairs, we plot the first $10^6$ pairs only.

**Definition 3:** Let us sort the pairs of nodes in the mesh in decreasing order of NSP. We define the mesh pair rank exponent $P_m$ to be the slope of the plot of the NSP versus the rank of the mesh node pairs in log-log scale.

**D. Mesh NSP Exponent $N_m$**

In this section, we study the distribution of NSP in the mesh. We define the frequency $f_{nm}$ of a NSP $n_m$ as the number of pairs with NSP of $n_m$ in the mesh. We plot the $(f_{nm}, n_m)$ pairs in log-log scale in Figure 9. In these plots, we exclude a small percentage of pairs of higher NSP that have lowest frequency, but still plot more than 99.9% of the total number of pairs. The solid lines are the result of the linear regression.

**Power-Law 4 (Mesh NSP Exponent):** The frequency $f_{nm}$ of a NSP between a pair of nodes in the mesh, $n_m$, is proportional to the NSP to the power of a constant $N_m$:

$$f_{nm} \propto n_{m}^{N_m}.$$  

**Definition 4:** We define the Mesh NSP exponent $N_m$ to be the slope of the plot of the frequency of the mesh NSP versus the mesh NSP in log-log scale.

**VII. Forest Topology Analysis**

In this section, we study the topology properties concerning the forest in the Gnutella network. In Table III, we present some basic statistics about the forest in our instances. In Table III, $p(t)$ represents the percentage of nodes in the forest.

**A. Tree Depth Distribution**

We define the probability $p(t_d)$ of a tree depth $t_d$ as the percentage of trees in the forest with depth $t_d$. Figure 10 describes the tree depth distribution.
From Figure 10, we notice that more than 56% of trees is simply composed of leaves directly connected to their corresponding root and more than 27% of trees have depth 2. In addition, less than 4% of trees have depth larger than 3.

B. Tree Rank Distribution

In this section, we study the size of each tree, which is defined as the sum of the vertices composing the tree plus the root. We sort the trees in decreasing tree size \(s_t\) and define tree rank \(r_t\) as the index of the tree in the sequence. We plot the \((s_t, r_t)\) pairs in Figure 11, applying log-scale only on the \(-y\)-axis. The solid lines are given by linear regression.

The plots of Figure 11 match the linear regression line. The ACC is 0.95621 for the 091505 instance and 0.95465 for the 021106 instance. Consequently, we infer the following empirical law and definition.

**Empirical Law 1:** The size \(s_t\) of a tree \(t\), is proportional to an exponential function with exponent being the product of the rank of the tree \(r_t\) and a constant \(T\):

\[
s_t \propto \exp(T r_t)\,.
\]

**Definition 5.** Let us sort the trees of a graph in decreasing order of size. We define \(T\) to be the slope of the plot of the sizes of trees versus the rank of the trees with log-scale applied on the sizes of trees.

This empirical law provides the formula on the sizes of trees in a sequence of trees.

VIII. Discussion

The regularity observed in our instances of the Gnutella network between September 2005 and February 2006 (including but not restricted to the two instances specifically discussed in this paper) is unlikely to be a coincidence. We could reasonably conjecture that our laws might continue to hold, at least for the near future.

Our work can facilitate the generation of realistic topologies of P2P networks, specially those which employ a hybrid and highly dynamic architecture like the Gnutella network. As an overview, we list the following guidelines for creating P2P network topologies. First, a small percentage of the nodes (15.4% or 10.0%) belong to the mesh. Second, the degree distribution of the mesh is skewed following our power-law 1 and 2, and the NSP distribution of the mesh is skewed following our power-law 3 and 4. Third, a large percentage of the nodes (84.6% or 90.0%) belong to the forest. Fourth, more than 56% of the trees have depth one, less than 4% of the trees have depth larger than 3, and the maximum depth is 7 or 10. Fifth, the size distribution of the trees is skewed following our empirical law 1. As a final step, we merge the generated mesh and the generated forest together to get the P2P network topology. If we finetune the parameters, we can get specific topologies that meet our needs.

IX. Conclusion

In this paper, we give a detailed review of the current Gnutella network topology as well as its on-going evolution. We present a novel two-layered approach to study the topology of the Gnutella network, analyzing it through the mesh perspective and the forest perspective respectively. Furthermore, we provide detailed topology properties as well as additional power-laws and the empirical law concerning these two layers. Using the two-layered approach and laws proposed, we can generate realistic topologies easily.

REFERENCES


