2008 Seventh International Conference on Grid and Cooperative Computing

# Scale-free Overlay Structures for Unstructured Peer-to-Peer Networks*

Jing Qi, Jiguo Yu †
School of Computer Science, Qufu Normal University
Ri-zhao, Shandong, 276826
jingqisd@126.com; jiguoyu@sina.com

## Abstract

*In unstructured peer-to-peer networks, the overlay topology among peers is a crucial factor in addition to the peer/data organization and search. The scale-free networks generated by the preferential attachment (PA) model have been widely described in previous researches. However, the PA model is incomplete to account for the robustness of real network. In this paper, the HLD and HHD models of scale-free overlay structures for unstructured peer-to-peer networks are proposed. With the identical degree distribution and network size in the models, we discuss the structural robustness and fragility as well as the dynamic changes of load intensity when nodes are successively removed under various attack strategies. Experimental results show that these two models are better than PA model in robustness and load-intensity.*

## 1. Introduction

In decentralized peer-to-peer networks, the overlay topology among peers is a crucial component in addition to the peer/data organization and search. Our work is related to peer-to-peer (P2P) network protocol design and topological analysis. Previous works on P2P network protocols can be classified into $centralized$ and $decentralized$. As centralized P2P protocols (e.g. Napster[15]) proved to be unscalable, the majority of the P2P research has focused on decentralized schemes. The decentralized P2P schemes can be further classified into sub-categories: $structured$, $unstructured$, and $hybrid$.

In the structured P2P networks, data/file contents of peers are organized based on a keying mechanism that can work in a distributed manner (e.g. CAN[16], Chord[19], Kademlia[14]). The keying mechanism typically maps the peers (or their contents) to a logical search space, which is then leveraged for performing efficient searches. Another positive side of the structured schemes is the guarantee of finding rare items in a timely manner. However, the cost comes from complexity of maintaining the consistency of mapping the peers to the logical search space, which typically causes considerable amount of control traffic (e.g. join/leave messaging) for highly dynamic P2P environments. Due to their capability of locating rare items, structured approaches have been very well suited to a wide-range of various applications[4],[6],[11].

In contrast to the structured schemes, unstructured P2P networks do not include a strict organization of peers or their contents. Since there is no particular keying or organization of the contents, the search techniques are typically based on flooding. Thus, the searches may take very long time for rare items, though popular items can be found very fast due to possible leveraging of locality of reference[12][20], and caching/replication[5][13].

To balance the trade-offs between structured and unstructured schemes, hybrid approaches[21] have attempted attaining a middle-ground between the costly maintenance of global peer/data keying of structured schemes and the high cost searches of unstructured schemes. Typically, hybrid schemes include a localized data/peer keying (e.g. DHT among neighbors instead of among all nodes) to achieve faster discovery of rare items, and a probabilistic search among the partially organized peers.

Our work is more applicable to unstructured P2P networks. Since we propose to use scale-free topologies in constructing the unstructured overlay P2P topology, we survey the scale-free network topologies in the following.

## 2. Related Work

Recent researches show that many natural and artificial systems such as the Internet[8], World Wide Web[1], scientific collaboration network[2], and e-mail network[7] have

*The work is supported by NNSF of China for contract 10471078, RFDP of Higher Education for contract 20040422004, Promotional Foundation for Middle-aged or Young Scientists of Shandong Province for contract (2005BS01016), EDRP of Shandong Province for contract(J07WH05), DRF and RF of QFNU for contract (XJ0609).

† the corresponding author

978-0-7695-3449-7/08 $25.00 © 2008 IEEE
DOI 10.1109/GCC.2008.34

power-law degree distributions. These systems are commonly known as power-law or scale-free networks since their degree distributions are not a function of the number of network nodes, but follow power-law distributions over many orders of magnitude. This phenomenon has been represented by the probability of having nodes with $k$ degrees as $P(k) \sim k^{-\gamma}$ where $\gamma$ is usually between 2 and 3 [3].

Scale-free networks have many interesting properties such as high tolerance to random errors and attacks, yet low tolerance to attacks targeted to "hub" nodes[9]. It has been also reported that the attack strategy based on betweenness centrality much harms network connectivity[10]. These results are obtained in the scale-free networks created through the preferential attachment rule[3]. As shown in recent papers[17][18], the algorithms depending on preferential attachment to construct the scale-free networks significantly vary. Our work lies in models of nodes preferably connecting based on degrees as well as preferential attachment rule.

The rest of the paper is organized as follows: we propose the network models in section 3, analyze some topologies indices and attack strategies in section 4, present our simulation results in section 5. Finally, we conclude the work.

## 3. Network Model

We discuss three models of scale-free network which generate different topological characteristics. First we construct a scale-free network with the preferential attachment rule, which is a basic growth algorithm to produce a network with a power-law degree distribution: The model evolves by one node at a time and the new node is connected to $m$ different existing nodes with probability proportional to their degree, i.e., $P_i = \frac{k_i}{\sum_j k_j}$ where $k_i$ is the degree of the node $i$. The links are regarded as bidirectional links. The special case of procedure is that the minimum degree $m$ is two, since we cannot make cycle structures in the case of $m < 2$. The algorithm for the PA model is presented in Figure 1. The algorithm assumes that the user has already created a network with $m + 1$ fully connected nodes. The function $ADDEDGE(i, j)$ creates an undirected edge between nodes $i$ and $j$. $RANDOM(i, j)$ creates a random integer $x$ such that $i \leq x \leq j$. $fRANDOM()$ creates a real-valued random number in $[0, 1]$. $Adj[i]$ means all the nodes are connected to node $i$. The variable $k_i$ stores the degree of the node $i$. $k_{total}$ stores the total number of degrees in the networks.

**Figure 1.** Pseudocode of the PA Model.

```
for i = m + 2 to N do
  for j = 1 to m do
    repeat
      try ← true
      node ← RANDOM(1, i-1); Rnd ← fRANDOM()
      if node ∉ Adj[i] AND Rnd< k_node / k_total then
        ADDEDGE(i, node)
        try ← false
      end if
    until try ← false
  end for
end for
```

Next, we change the network with fixing the degree distribution $P(k)$ of the preferential attachment model. One of the models, namely HLD model, generates the network where high-degree nodes preferably connect to the low-degree nodes. First, based on the degree distribution $P(k)$, all nodes are labeled and arranged in a queue as $n_1, n_2, ..., n_N$ in descending order of their degree. The network is produced by the algorithm iterating the following procedures: selecting node $n_i$ with the highest degree $k_i$ from the head of the queue, linking edges between $n_i$ and $k_i$ nodes picked up randomly from the queue $(n_{i+1}, ..., n_N)$, and removing nodes fully connected from the queue. The algorithm of the HLD model is presented in Figure 2. The algorithm assumes that the user has already sorted the nodes in descending order of their degrees.

**Figure 2.** Pseudocode of the HLD Model.

```
for i = 1 to N do
  for j = i + 1 to N do
    repeat
      try ← true
      node ← RANDOM(i+1, j); Rnd ← fRANDOM()
      if node ∉ Adj[i] AND Rnd< k_node / k_total then
        ADDEDGE(i, node)
        try ← false
      end if
    until try ← false
  end for
end for
```

The HHD model generates the network where high-degree nodes preferably connect to the other high-degree nodes. The algorithm is similar to the HLD model with replacing the procedure: linking edges between $n_i$ and $k_i$ nodes picked up randomly from the queue depending on their weights $w_{i+1}, ..., w_N (w_a = k_a^p, p > 1)$.

## 4. Analysis

### 4.1. Topological Parameters

We quantitatively evaluate the structural properties of these models with some topological indices. All the networks in this study have 1000 nodes and 2000 edges. In case a slight topological difference may change the results of analysis, the results shown in this paper are averaged by 100 trials.

The average path length $\overline{L}$ of a network is given by

$$\overline{L} = \frac{2\sum_{i<j} L_{ij}}{N(N-1)} \tag{1}$$

where $L_{ij}$ is the shortest path length between node $i$ and node $j$, and $N$ is the number of nodes.

The clustering characteristic of node $i$ is evaluated by the clustering coefficient $C_i$ expressed as

$$C_i = \frac{2E_i}{k_i(k_i-1)} \tag{2}$$

where $k_i$ is the degree of node $i$ and $E_i$ is the number of edges that exist between these $k_i$ nodes. The average clustering coefficient of the network is given by $\overline{C} = \frac{\sum_i C_i}{N}$.

The betweenness centrality $B_i$ of node $i$ is given by

$$B_i = \sum_{s<t} \frac{\sigma_{st}^i}{\sigma_{st}} \tag{3}$$

where $\sigma_{st}$ is the total number of the shortest paths from node $s$ to node $t$ and $\sigma_{st}^i$ is the number of the shortest paths from $s$ to $t$ passing through node $i$. Nodes of high betweenness centrality have a key role to shorten the length of many paths between nodes in the network, and tend to have high load intensity. The average betweenness centrality of the network is given by $\overline{B} = \frac{\sum_i B_i}{N}$.

The distribution of betweenness centrality is an important factor to estimate the efficiency of traffic load on the network, and we thus evaluate the betweenness deviation expressed as

$$\delta_B = \sqrt{\frac{\sum_i (B_i - \overline{B})^2}{N}} \tag{4}$$

The low betweenness deviation means that load on the network is efficiently distributed, whereas the high betweenness deviation means that load is concentrated in a part of the network.

**Table 1: Topology Parameters**

| model type | $\overline{L}$ | $\overline{C}$ | $\overline{B}$ | $\delta_B$ |
|---|---|---|---|---|
| PA model | 4.05 | 0.03 | 1521 | 6543 |
| HLD model | 4.34 | 0.01 | 1669 | 5150 |
| HHD model | 4.89 | 0.02 | 1941 | 6133 |

Table 1 shows the topological parameters for the three models in the initial state. As shown in Table 1, the betweenness deviation of the HLD model is significantly smaller than that of the PA model. This means that the load on the HLD model is efficiently distributed. Although the HHD model has the same power-law degree distribution, it has the long average path length. In addition, the average betweenness centrality of the HHD model is quite high. This indicates that there are many bottlenecks in the HHD model in the initial state.

### 4.2. Attack Strategies

The robustness of networks has been previously analyzed under the strategies of node attack especially based on the node's degree. In addition to the degree-based attack, the betweenness-based attack also belongs to the harmful strategy.

In this paper, we classify the attack strategies depending on whether they are based on degree or betweenness centrality and whether the attack is determined statically or dynamically. The static degree-based attack is the strategy that the target node is selected one by one in descending order of its initial degree. The dynamic degree-based attack is the strategy that the target node is dynamically selected by searching a node with the highest degree at every removal step. In the same way, the static betweenness-based attack and the dynamic betweenness-based attack target the node with highest betweenness centrality. When a node is removed from networks, all edges connecting the node are also removed.

## 5. Experimental Results

### 5.1. Network Robustness

The fraction of the removed node is represented by $f = \frac{rm}{N}$, where $rm$ is the number of removed nodes and $N$ is the number of nodes in the initial connected network. The fraction of nodes contained in the largest connected component out of the $N$ nodes is represented by $lc$. To determine the critical threshold $f_c$ at which a network is disrupted, we illustrate the $lc$ and $\overline{L}$ of the largest connected component in the function of the $f$ in Figure 3 and Figure 4.

As nodes are removed, the network breaks into isolated clusters gradually as well as $lc$ decreases. Meanwhile, $\overline{L}$ increases and peaks just before the network is disrupted. After the network breaks into isolated clusters, $\overline{L}$ decreases rapidly because the largest connected component decreases. In this study, the critical threshold $f_c$ is determined by the fraction of removed nodes from the network when $\overline{L}$ is the maximum. Large $f_c$ indicates that the network is robust in terms of connectivity. From our results, we find that the

HHD model is more robust than the PA model and the HLD model under all the attack strategies.

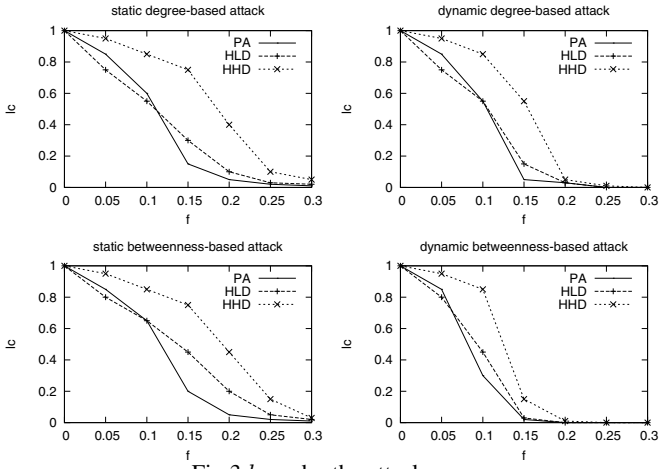


Fig.3 $lc$ under the attacks

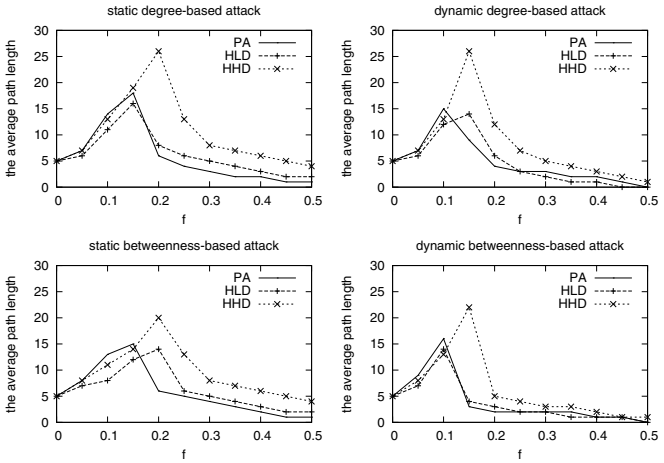


Fig.4 $\overline{L}$ under the attacks

## 5.2. Network Load

To analyze the dynamic change of the amount of load and its distribution, we monitor the average node betweenness centrality $\overline{B}$ and the betweenness deviation $\delta_B$ of the largest connected component in Figure 5 and Figure 6. $f_B$ represents the fraction of node removal when the average betweenness centrality is maximum.

Under these attack strategies, $\overline{B}$ for the PA model and the HHD model increases rapidly as $f$ increases, and peaks at $f_B$ before the critical threshold $f_c$. These results indicate that the break-downs of nodes make the amount of load heavier when these networks are attacked. After these networks break into isolated clusters, the average betweenness centrality $\overline{B}$ decreases rapidly as well as the average path length $\overline{L}$. In addition, the betweenness deviation $\delta_B$ once drops by removal of a few nodes. As nodes are removed, $\delta_B$ increases again, and peaks near $f_B$, indicating that the distribution of betweenness becomes heterogeneous by successive attacks. The increase in the heterogeneity of betweenness distribution indicates that the excessive load is concentrated on a few nodes; therefore, the PA and the HHD models tend to cause the load congestion against attacks. For the HLD model, both $\overline{B}$ and $\delta_B$ do not increase very much when the nodes are removed successively except under the dynamic betweenness-based attack strategy. Even under the dynamic betweenness-based attack strategy, the values of $\overline{B}$ and $\delta_b$ for the HLD model are much smaller than those of the other scale-free models. Since the HLD model can maintain the homogeneity of betweenness distribution against attacks, it is a highly load-tolerant structure.

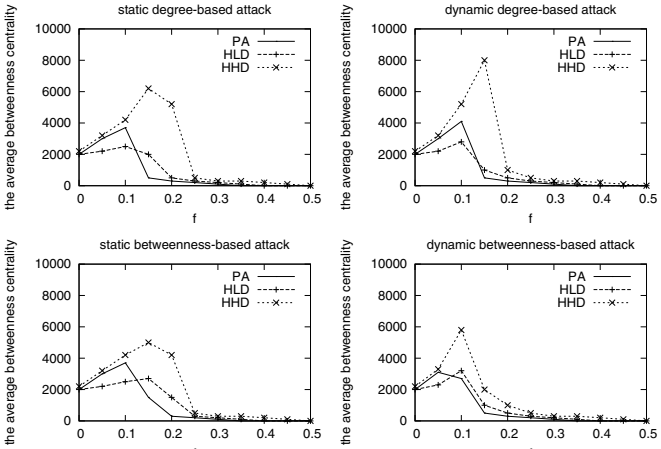


Fig.5 $\overline{B}$ under the attacks

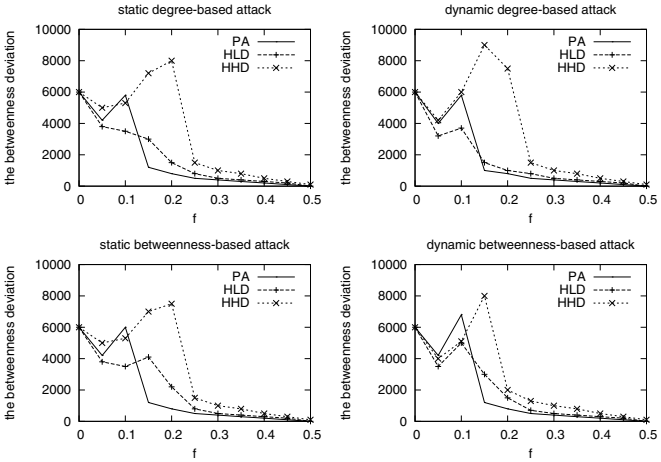

Fig.6 $\delta_B$ under the attacks

## 6. Conclusion

The preferential attachment model has been widely used to study the basic functionalities of real networks. However, it is clear that this model is incomplete to account for the robustness of real networks. We propose the new models of scale-free overlay structures for unstructured peer-to-peer networks in this paper. We study the connectivity robustness and load tolerance under the attack strategies in these models. The results indicate that the new models are structurally robust against attacks and highly load-tolerant. We will apply the new models into the real networks in the future.

## References


[1] R. Albert, H. Jeong, Diameter of the world wide web, *Nature, 1999.*

[2] A. L. Barabasi, Evolution of the social network of scientific collaborations, *Physica A, 2002.*

[3] A. L. Barabasi and R. Albert, Emergence of scaling in random networks, *Science, 1999.*

[4] M. Castro, P. Druschel, Scalable application-level anycast for highly dynamic groups, *Networked Group Communications(NGC), 2003.*

[5] E. Cohen and S. Shenker, Replication strategies in untructured peer-to-peer networks, *ACM SIGCOMM, 2002.*

[6] R. Cox, A. Muthitacharoen, and R. Morris, Serving dns using a peer-to-peer lookup service, *Workshop on Peer-to-Peer Systems (IPTPS), 2002.*

[7] H. Ebel, M. I. Mielsch, and S. Bornholdt, Scale-free topology of e-mail networks, *Physical Review E, 2002.*

[8] M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships of the internet topology, *Computer Communications, 1999.*

[9] Hasan Guclu, Murat Yuksel, Scale-Free overlay topologies with hard cutoffs for unstructured peer-to-peer networks, *Distributed Computing Systems(ICDCS), 2007.*

[10] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, Attack vulnerability of complex networks, *Physical Review E, 2002.*

[11] D. Kato, Gisp: Global information sharing protocol-a distributed index for peer-to-peer systems, *The Second International Conference on Peer-to-Peer Computing(P2P), 2002.*

[12] Y. Liu, L. Xiao, X. Liu, L. M. Ni, and X. Zhang, Location awarenes in unstructured peer-to-peer systems, *IEEE Transactions on Parallel and Distributed Systems, 2005.*

[13] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker, Search and replication strategies in unstructured peer-to-peer networks, *ACM International Conference on Supercomputing(ICS), 2002.*

[14] P. Maymounkov and D. Mazieres, Kademlia: A peer-to-peer information system based on the xor metric, *Workshop on Peer-to-Peer Systems (IPTPS), 2002.*

[15] Napster, available at http://www.napster.com.

[16] S. Rantasamy, P. Francis, M. Handley, R. Karp, and S. Schenker, A scalable content-addressable network, *ACM SIGCOMM, 2001.*

[17] H. Wouhaybi Rita, T. Campbell Andrew, Phenix: Supporting resilient low-diameter peer-to-peer topologies, *IEEE INFOCOM, 2004.*

[18] M. Sasabe, N. Wakamiya, M. Murata, LLR: A construction scheme of a low-diameter, location-aware, and resilient P2P network, *Workshop on Mobility, Collaborative Working, and Emerging Applications(MobCops), 2006.*

[19] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, Chord: A scalable peer-to-peer lookup service for internet applications, *Computer Communications, 2001.*

[20] Chi-Jen Wu, De-Kai Liu,and Ren-Hung Hwang, A location-aware peer-to-peer overlay network, *International Journal of Communication Systems, 2007.*

[21] Chao Xie, Guihai Chen, Art Vandenberg, Yi Pan, Analysis of hybrid P2P overlay network topology. *Computer Communications, 2008.*