Advances in Complex Systems, Vol. 14, No. 4 (2011) 611–633

DOI: 10.1142/S0219525911003141

World Scientific
www.worldscientific.com

ON COVERAGE BOUNDS OF UNSTRUCTURED PEER-TO-PEER NETWORKS

JOYDEEP CHANDRA* and NILOY GANGULY†

Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur, India
**joydeep@cse.iitkgp.ernet.in*
†*niloy@cse.iitkgp.ernet.in*

Received 15 February 2011
Revised 26 May 2011

In this paper, we develop methods to estimate the network coverage of a *TTL*-bound query packet undergoing flooding on an unstructured p2p network. The estimation based on the degree distribution of the networks, reveals that the presence of certain cycle-forming edges, that we name as cross and back edges, reduces the coverage of the peers in p2p networks and also generate a large number of redundant messages, thus wasting precious bandwidth. We therefore develop models to estimate the back/cross edge probabilities and the network coverage of the peers in the presence of these back and cross edges. Extensive simulation is done on random, power-law and Gnutella networks to verify the correctness of the model. The results highlight the fact that for real p2p networks, which are large but finite, the percentage of back/cross edges can increase enormously with increasing distance from a source node, thus leading to huge traffic redundancy.

Keywords: Peer-to-peer networks; network coverage; network performance; overlay characteristics; message redundancy.

1. Introduction

Many unstructured peer-to-peer (p2p) networks like Gnutella, FreeHaven and Kazaa use broadcasting as their query and search mechanism [12]. The query and search performance of these p2p networks are directly proportional to the network coverage of the peers achieved through broadcasting. Since, broadcasting or flooding generates huge amount of traffic [3, 21], to control the volume of traffic, most networks limit the number of hops of a broadcast message to a particular Time-to-Live (*TTL*) value, naturally sacrificing coverage hence search efficiency. For example, dynamic querying [1] in Gnutella uses *TTL*(2) (numeric value inside parenthesis represents the number of hops to search with) flooding for most searches. Apart from the p2p networks, in context to which the importance of network coverage has been discussed here, the impact of network coverage extends to many other important applications like online social networks, online recommendation systems and also online advertising systems. There have been some recent works on

improving network coverage for increasing search efficiency through better topology management. In [8, 28] the authors have proposed cycle-free topology generation mechanisms for improving network coverage and message redundancy in networks that use $TTL(2)$ flooding; however, the works are not supported by proper analysis of the coverage bounds. Thus proper analysis of the coverage bounds of the peers in the network is required to understand the effectiveness of the strategies; this paper is directed towards fulfilling that goal.

Topological properties of large graphs have been a much investigated topic in classical physics [9–11, 26], in context of which Newman *et al.* [17] studied the neighborhood properties of nodes in large random networks with arbitrary degree distributions using a generating function formalism. We have used these concepts of Newman *et al.* in our earlier paper [7] to derive a basic model of $TTL(2)$ coverage bounds of the peers in p2p networks with a given degree distribution. The topology of these networks was assumed to be *random*, where a particular topology can be selected uniformly from all the possible topologies with such a given degree distribution. We have also identified that the basic model makes a simplified assumption that the underlying topology is tree-like; in contrast real networks contain certain cycle forming edges, which we referred as back and cross edges. In a network, if a source node broadcasts a TTL based message, the cross and back edges of the source node are those edges, through which a transmitted non-duplicate message from one end of the edge will produce a duplicate message at the receiving node. In the process of broadcasting the TTL-based message of a source, the probability of encountering these cross and back edges are termed as the cross and back edge probabilities, respectively, of the source node.

In Ref. 7, we introduced the concept of cross/back edges and derived the $TTL(2)$ coverage of the peers in these networks, assuming a common back and cross edge probability for all the nodes in the network. No efforts were made to derive the back and cross edge probability of the nodes in the network. However, from the simulations (the details of which is discussed later) we observed that the number of back/cross edges of a node depend on its current degree. Hence, based on this observation, in this paper we derive models to estimate the back/cross edge probabilities of the peers, with respect to their degrees, for random networks when certain network properties like the degree distribution and clustering coefficients are known. The proposed refinement thus derives the $TTL(2)$ network coverage of the peers with respect to their current degree. The accuracy of the models is validated on several types of networks, like Erdős–Rényi networks, scale-free networks, random clustered networks [19] and also on a simulated Gnutella-like network using extensive simulations. We further generalize the concept of back and cross edges for any TTL values and derive the $TTL(n)$ network coverage of the peers in the network. The results indicate that the probabilities of occurrence of these edges increase enormously with increasing distance from the source nodes, which can result in huge traffic redundancy, thus questioning the effectiveness of larger TTL based search. Although the models have been derived and validated for p2p

based applications, however they are applicable for various other network-based information spreading or biological applications that assumes random network topologies.

The organization of the paper is as follows: in Sec. 2, we provide a brief background of the topology formation and search mechanism in unstructured p2p networks. In Sec. 3, we provide a brief review of the basic model and state its limitations, in Sec. 4, we derive a refined model of estimating $TTL(2)$ network coverage for given cross/back edge probabilities. We derive the back/cross edge probabilities for purely random as well as clustered random networks with arbitrary distribution in Sec. 5; in Sec. 6, we derive the $TTL(n)$ network coverage and generalize the back/cross edge probabilities for any distance from a source node. Finally, we draw conclusions in Sec. 7.

2. Overview of Unstructured p2p Networks

We here provide a brief outline of the characteristics of unstructured p2p networks. In these networks, the peers or the nodes that join a particular p2p network form a community among themselves. The peers store certain files, which are shared over the network community. Peers willing to obtain a file from any remote peer initially issues a query for that file; the peers storing the intended file are searched over the network and if the file exists, it is downloaded from the peers where it resides. Although the precise characteristics of the various unstructured p2p networks differ slightly, however, the topological structure and the search mechanism in these networks are largely the same.

Popular unstructured p2p systems like Gnutella[a] and Kazaa use a super-peer based architecture, where a set of few resourceful (like having high storage space, high bandwidth etc.) peers act as super-peers or ultra-peers, whereas, rest of the peers are termed as leaf peers. Each peer runs the p2p application software that we term as *servant*, which performs all the basic p2p tasks like bootstrapping, communicating with other peers and file retrieval. Apart from these basic tasks, the ultra-peers handle the search and indexing mechanism in the network. Although the exact values of the number of ultra and leaf neighbors depend on the specific servant that is being used by the peers but for the case of Limewire servant, which is a very popular Gnutella-based servant [23], each ultra-peer connects to a maximum of 30 leaf peers and maintains a hashed index of the files that are present in the leaf peers to which it is directly connected. Each ultra-peer is connected to a maximum of 32 other ultra-peers; the ultra-peers are responsible for searching the files requested by peers. Each leaf peer connects to 3–4 ultra-peers; however, the leaf peers are not connected to each other. Thus the majority of the traffic moves on the ultra-peer level. Early measurements studies have shown that there

[a]From version 0.6 onwards, the Gnutella networks switched to a two layer super-peer based architecture; previous versions used a pure decentralized model, where all the nodes are considered equal.

are approximately 100–200 K live Gnutella peers in the network at any time out of which nearly 15–16% are ultra-peers [15].

Some peers maintain information about other remote peers along with their distance, in terms of the number of hops from it, in a cache called GWebCache. The information about these remote peers is propagated through their neighbors; newly joined peers obtain the information about these remote peers from some of these GWebCaches and attempts to connect to them, randomly. This protocol determines the topological nature of the system. The topological properties of these unstructured p2p systems have been studied in details [16, 20, 25] and it has been observed that the degree distribution of the peers follows a power-law distribution. This is because, due to higher connectivity, the information about the high degree peers disseminates to more number of nodes as compared to the low degree peers; hence newly joined peers are more likely to connect to a high degree peer as compared to a low degree one (preferential attachment), thus forming a power-law degree distribution of the peers. Further, it has also been observed that the average clustering coefficient in Gnutella is higher (0.02) as compared to Erdős–Rényi graphs of the same size (0.002); however, the exact reasons for this high clustering is not yet known. Hence, keeping in view the clustering properties of the nodes, we derive our models for two kind of random network topologies, (a) uncorrelated random networks with given degree distribution generated using configuration model and (b) clustered random networks with given degree distribution and average clustering coefficients of each node degree, generated using Serrano model [22], that generates random networks with given clustering properties.

The query propagation mechanism in unstructured p2p networks is based on flood-based mechanisms. For example, the query propagation mechanism from Gnutella version 0.6 onwards are done using a technique known as *Dynamic Querying* mechanism [1]. In this method, the ultra-peers use a controlled broadcasting means to propagate the queries to other peers. Initially, a query is sent by an ultra-peer to its neighbors with TTL value of 1. If the total number of search results returned by the neighbors is greater than 150, the query propagation is stopped, otherwise the query is propagated with a TTL value of 2 to a subset of neighbors. A TTL value of 3 is only used for very rare searches when the number of results returned by $TTL(1)$ and $TTL(2)$ broadcast is less than 150. Modern Gnutella servants never propagate a query with TTL value greater than 4; Kazaa, also uses a similar TTL value. Thus, the search performance of the peers is directly proportional to their network coverage, i.e. the number of unique peers to which the TTL bound queries can reach. Hence, we initially derive the network coverage of the peers for up to TTL values of 2 and then later generalize the same for any TTL.

We next provide a brief review of the basic model, but prior to that a list of main notations that will be used throughout the paper is summarized in Table 1 for ready reference.

Table 1. Notations.

$TTL(r)$	Query with $TTL = r$
N	Total number of peers in the network
p_i	Probability that a random node in the network is of degree i
$\langle z \rangle$	Average degree of the peers in the network
$\langle z^2 \rangle$	Second moment of the degree of the peers
κ_k	Cross edge probability at level 1 of peers with degree k
b_k	Back edge probability at level 1 of peers with degree k
$\kappa_k^{(c)}$	Cross edge probability at level 1 in a clustered random network of peers with degree k
$b_k^{(c)}$	Back edge probability at level 1 in a clustered random network of peers with degree k
$\kappa_k(l)$	Cross edge probability at level l of the peers of degree k
$b_k(l)$	Back edge probability at level l of the peers of degree k
$a_k(l)$	Total number of neighbors of a peer, of degree k, up to level l

3. The Basic Model and Its Limitations

In this section we present a brief review of the basic model for deriving the $TTL(2)$ coverage of the peers in a random network with given degree distribution and discuss the limitations of this model. The degree of the nodes are chosen randomly based on the specified distribution; the network topology is assumed to be chosen randomly from all the possible topologies with that given degree distribution.

The coverage of a peer for a $TTL(2)$ broadcast is actually the sum of its first and second neighbors. Newman *et al.* [17] derived models for the distribution of the number of first and second neighbors of a node in a large graph, where the number of nodes, $N \to \infty$. Suppose, in a large network with N nodes (N is large), let p_k denote the probability of any random node in a network having k first neighbors. Such an uncorrelated random network can be easily generated using the configuration model [6]. The first neighbor distribution of the nodes in the network can be represented using a generating function [27] as,

$$G_0(x) = p_0 + p_1 x + p_2 x^2 + p_3 x^3 \ldots, \tag{1}$$

where the coefficient of x^i in $G_0(x)$ gives the probability that any random node in the network will have degree i. The average number of neighbors of a node is given by,

$$\langle z \rangle = 1 \cdot p_1 + 2 \cdot p_2 + 3 \cdot p_3 + \cdots = G_0'(1). \tag{2}$$

On traversing a random edge, the probability of reaching a node with $k-1$ outgoing edges is proportional to both its degree k and also the probability of selecting the node of degree k, which is p_k. Thus, the probability of reaching a node with $k-1$ outgoing edges is given as $\frac{kp_k}{\sum_j jp_j} = \frac{kp_k}{\langle z \rangle}$. The generating function for the distribution

of the remaining outgoing edges of a node reached by following a random edge can thus be represented as,

$$G_1(x) = \frac{1}{\langle z \rangle} \cdot \left(\sum k p_k x^{k-1} \right) = \frac{G_0'(x)}{G_0'(1)}. \tag{3}$$

Suppose, we want to find the number of second neighbors of a node, P. The distribution of the number of outgoing edges from k neighbors of a random node with degree k is given by

$$S_k(x) = [G_1(x)]^k. \tag{4}$$

Thus, if each of the outgoing edges from the k neighbors leads to a unique node (i.e. no cycles are formed), then $S_k(x)$ represents the distribution of the number of second neighbors of a node with degree k and the distribution of the number of second neighbors of any random node P, is given by,

$$S(x) = \sum_k p_k [G_1(x)]^k = G_0(G_1(x)). \tag{5}$$

If $\langle z_2 \rangle$ denotes the average number of second neighbors of a node, then

$$\langle z_2 \rangle = \left[\frac{d}{dx} G_0(G_1(x)) \right]_{x=1} = G_0''(1). \tag{6}$$

The total network coverage of a peer in p2p networks that use $TTL(2)$ flooding scheme is the sum of its number of first and second neighbors. Hence the distribution of the total node coverage of a peer P that deploys a $TTL(2)$ flooding mechanism is represented by the generating function $C(x)$ as,

$$C(x) = \sum_k p_k x^k S_k(x) = \sum_k p_k [x G_1(x)]^k = G_0(x G_1(x)). \tag{7}$$

Using these expressions, we can obtain the expected $TTL(2)$ coverage, $\langle c \rangle$ of a peer which is given as,

$$\langle c \rangle = C'(1) = \left[\frac{d}{dx} G_0(x G_1(x)) \right]_{x=1} = G_0'(1) + G_0''(1) = \langle z \rangle + \langle z_2 \rangle. \tag{8}$$

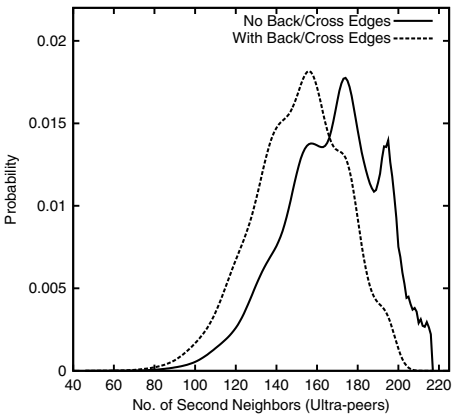
Limitations of the Basic Model. The expression in Eq. (7) provides correct reachability distributions only when the nodes reached from a source node, using $TTL(2)$ flooding, do not form any short length cycles among themselves, i.e. it actually provides a maximum $TTL(2)$ coverage bound of a node for a given degree distribution. This is because, the expression $S_k(x) = [G_1(x)]^k$ in Eq. (4) gives the distribution of the number of excess outgoing edges from the first neighbors of a k-degree node, P. Assuming that each outgoing edge from the neighbors of P leads to a unique node, i.e. a node that has not been reached through any other path, $S_k(x)$ represents the distribution of the number of second neighbors of P. But,

for many real cases, this condition fails to hold; for example, in p2p systems that behave like social networks, the nodes inherently form many short length cycles. The cycles, caused by edges that affect the coverage of the nodes are referred to as *cross* and *back* edges (see Fig. 3(a)), which are defined as follows.

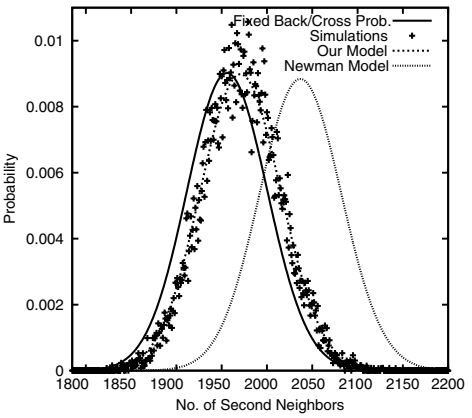
Definition 1. To define cross and back edges of a source node (say P), we perform a breadth first traversal from P. A node P_i is said to be at level i with respect to P if P_i is reached using a minimum of i hops from P. Considering the root P to be at level 0, a cross edge at level i of node P is formed, if an edge from a node at level i of P, connects to another node at the same level (in Fig. 3(a) edges P_1P_2, P_3P_4 are cross edges of P at level 1 and XY is a cross edge of P at level 2) or to a node at level $i-1$ except its parent (in Fig. 3(a), XP_2 is a cross edge of P at level 2). To define back edges of node P at level i, we traverse the edges from the nodes at level i, in any random order, to reach the nodes at level $i+1$; an edge is defined as a level i back edge of node P, if the edge connects to a node at level $i+1$, which has already been reached through a different node (in Fig. 3(a), edge P_2X is a level-1 back edge of P as P_1X has been traversed earlier than P_2X). Thus, edges are defined with respect to nodes; hence an edge can be a back-edge with respect to a node and cross edge with respect to other. Further, for a particular node, the same edge can be a back edge at a particular level i and cross edge at level $i+1$ (in Fig. 3(a), edge P_2X is a back edge of P at level 1 and a cross edge at level 2.

Every other edge will always lead to a new node that has not been explored earlier and will be termed as regular edge. For simplicity, we refer the back/cross edges at level 1 as simply back/cross edge,without repeatedly mentioning the term, level 1, whereas back/cross edges at level l ($l > 1$) are referred explicitly as back/cross edges at level l. In this paper, unless otherwise mentioned explicitly, the terms back edge and cross edge for a source node refer to back and cross edge at level 1 only. Certain salient features of the back/cross edges observed by simulating a prototype of Gnutella network and an Erdős–Rényi network are as follows:

(1) The presence of back and cross edges in a network reduces the average second neighbors of the nodes in the network. Figure 1(a) shows the effect of back/cross edges on the distribution of second neighbors of the Gnutella network with around 27 k nodes. Due to the presence of back/cross edges, the average number of second neighbors for nodes with degree 7 is 151, which would have been 168 otherwise, thus resulting in 10% drop in coverage at $TTL(2)$.
(2) The back/cross edge probabilities are dependent on the degree k of a source node and cannot be considered a single value for the entire network (an initial suggestion in Ref. 7). Figure 1(b) illustrates the situation. In this figure, we compare the distribution of second neighbors in a Erdős–Rényi network, of a node of degree 40, obtained from our model by considering fixed back/cross edge probability and degree dependent back/cross edge probability. The root mean squared error (RMSE) value derived for our refined model, that considers



(a) Back/cross edge effect



(b) Variation of average second neighbor with back/cross edge

Fig. 1. (a) shows the effect of back/cross edge on the second neighbor distribution of the peers in a Gnutella Network; the actual number of second neighbors is much less in the presence of back/cross edges as compared to the case when no back cross edge is present. (b) shows the accuracy of the calculated average second neighbors when back/cross edge probabilities are calculated as a function of the source degree as compared to the case when a fixed probability is used for the whole network. The network size is 30 k with connection probability $\hat{p} = 0.0017$. The graph compares the second neighbor distribution of nodes with degree $k = 40$, obtained from simulation results with 3 possible cases, (i) when a fixed mean back and cross edge probability, $b_k = 0.04$ and $\kappa_k = 0.001$, respectively, is used, (ii) when back/cross edge probability is calculated using our model and (iii) for the basic model when no cross/back edge is assumed.

degree dependent back/cross edge probabilities, with respect to the simulation results was found to be 0.0005, whereas the RMSE for fixed back/cross edge probabilities with respect to the simulation results was around 0.0013. The normalized RMSE (NRMSE) obtained by dividing the RMSE value with the difference between the maximum and minimum values of the probabilities of second neighbors is around 0.05 and 0.13, respectively, thus indicating that degree dependent back/cross edge probabilities produces a better fit as compared to the fixed back/cross edge probabilities.

Thus the task is to derive a rigorous model which will estimate the second neighbor distribution of a network taking into consideration the degree-dependent back/cross edge probabilities which is done in the next two sections.

4. Coverage Bound With Back and Cross Edges: Refined Model

In this section, we derive models for two-hop network coverage of a node in any random network when the degree distribution and the back/cross edge probability of the network is known. Thus unlike the basic model, where apart from the degree distribution, all other aspects of the network was random, in this case we consider networks, where the degree distribution and the cross/back edge probabilities of

the nodes in the network are given; however connectivity between the nodes is totally random. We later go on to derive back/cross edge probabilities for certain networks like uncorrelated random networks, Erdős–Rényi networks and random networks with given clustering coefficient. Further, the derived two-hop coverage model will be used later to derive the generalized coverage of the peers for any TTL values. We later validate the models using simulations on various network topologies like (a) Erdős–Rényi networks formed by connecting each pair of nodes in the network with a fixed probability, (b) uncorrelated power-law networks formed by distributing the degree of the peers according to a power-law ($p_k \sim k^{-\alpha}$, where α is a constant that varies from 2 to 3 for most real networks [4, 5]) and connecting the peers randomly using configuration model[b] [6] and (c) the ultra-peer level network topology in a simulated Gnutella-like network [8], so as to obtain an arbitrary degree distribution. We refer this topology as a *Simulated-Gnutella* topology. Apart from these we also validate our protocols on topologies that are generated based on real Gnutella snapshots obtained from Ref. 2 like, (d) an uncorrelated random network formed using the degree distribution of the peers in real Gnutella and connecting these peers using the configuration model (*Random-Gnutella*) and (e) a clustered random network based on the degree distribution and clustering coefficients for each node degree in the real Gnutella network, generated using an algorithm proposed by Serrano *et al.* [22] (*Serrano-Gnutella*).

Let us assume that a network with known degree distribution has N nodes and a random node P has degree k. Then as shown in Eq. (4), the number of outgoing edges from the first neighbors of node P follows a distribution that can be represented using a generating function as

$$S_k(x) = [G_1(x)]^k.$$

However, due to the presence of back or cross edges, the underlying independent assumption that all the outgoing edges results in distinct neighboring nodes does not hold. *We define the cross/back edge probability at level i of a node P with degree k (denoted as $\kappa_k(i)$ and $b_k(i)$ respectively) as the probability that a randomly picked outgoing edge from any of its ith hop neighbor is a cross/back edge.* However for level 1, we denote the cross and back edge probabilities for a source node with degree k ($\kappa_k(1)$ and $b_k(1)$, respectively) as κ_k and b_k, respectively. Then the probability that any randomly chosen outgoing edge from the first neighbors of P is a regular edge (i.e. neither a back nor a cross edge) is given by,

$$w_k = 1 - b_k - \kappa_k. \tag{9}$$

[b]Configuration Model: The degree distribution of the peers are used to generate the degree of each node, i.e. if the total number of nodes in the system is N, then the number of k-degree peers formed is Np_k; these peers were selected randomly and assigned k free stubs. This process is repeated for all possible values of k. The network is then formed by selecting two disconnected nodes randomly having at least one free stub each and connecting these stubs to form an edge. This process is repeated unless no free stub remains.

For a node P with degree k, and having t outgoing edges from its set of first neighbors, let γ denote the number of regular edges from the first neighbors of P, the probability of which is given as $\binom{t}{\gamma} w_k^{\gamma}(1-w_k)^{t-\gamma}x^{\gamma}$; thus the generating function for the distribution of the number of second neighbors of a node, P, can be represented as,

$$Q_{k,t}(x) = \sum_{\gamma \le t} \binom{t}{\gamma} w_k^{\gamma}(1-w_k)^{t-\gamma}x^{\gamma}. \tag{10}$$

According to Eq. (4), $S_k(x)$ denotes the generating function for the distribution of the remaining outgoing edges from the first neighbors of a node with degree k. Let $S_k(x)$ be represented as,

$$S_k(x) = s_{k,0} + s_{k,1}x + s_{k,2}x^2 + \cdots + s_{k,t}x^t + \cdots . \tag{11}$$

The coefficient of x^t in Eq. (11) represents the probability of t remaining outgoing edges from the first neighbors of a node with degree k. An outgoing edge leads to a unique second neighbor if the edge is a regular edge. Thus the distribution for the unique second neighbors of a node with k first neighbors is given by,

$$A_k(x) = \sum_t s_{k,t}Q_{k,t}(x). \tag{12}$$

The distribution of second neighbors for any random node in a network is,

$$\hat{S}(x) = \sum_{k'} p_{k'}A_{k'}(x). \tag{13}$$

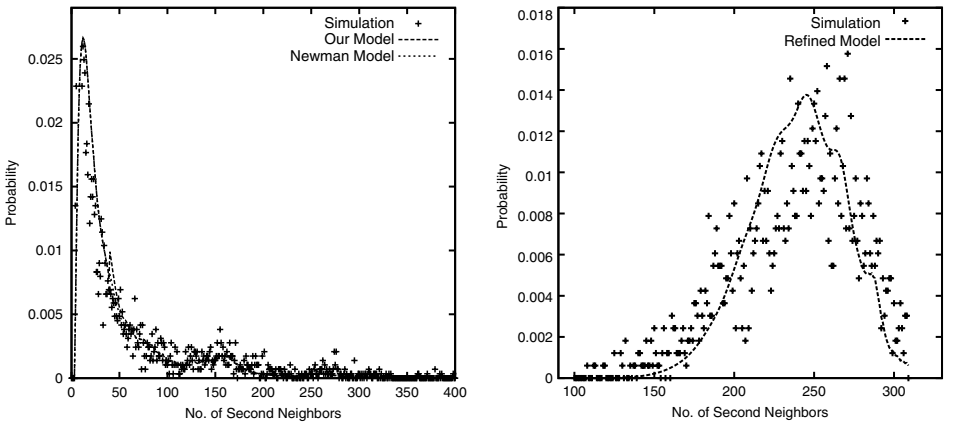
The distribution for total coverage of the network will be given by $\hat{C}(x) = \sum_k p_k x^k A_k(x)$. Thus the average $TTL(2)$ network coverage of a random source node in the presence of back/cross edges is given as,

$$\langle c \rangle = \hat{C}'(1). \tag{14}$$

We discuss next the simulation results that we have obtained in order to validate the above derived model.

Simulations. We have simulated and compared the second neighbor distribution of nodes derived from our model for three types of networks (a) Erdős–Rényi network, (b) uncorrelated power-law network and (c) simulated-Gnutella network. For all cases, the back and cross edge probabilities (b_k and κ_k, respectively) used in the model were derived using separate methodology that we discuss later.

The simulation result for the Poisson graphs with 30 K nodes, having an average degree of 51 is shown in Fig. 1(b). As can be seen in figure, the second neighbor distribution calculated using our model matches almost exactly with the simulation results. The RMSE value of our model results, with respect to the simulation results is 0.0005 (NRMSE = 0.05), whereas for the case of Newman model, the RMSE is around 0.0046 (NRMSE = 0.46), thus indicating that our model produces a much better fit as compared to the Newman model.



(a) Power-law (b) Random-Gnutella network

Fig. 2. Second neighbor distribution of a node with $k = 4$ first neighbors for power law network with $N = 30\,\mathrm{K}$ nodes, $\alpha = 2.78$, back probability $b_k = 0.0024$ and cross probability $\kappa_k = 0.0013$, and a Random-Gnutella network with $N \approx 27\,\mathrm{K}$ ultra nodes, $b_k = 0.0043$ and $\kappa_k = 0.0002$. The points show the simulation results, the heavy broken lines indicate the results of our refined model. (Figures replotted.)

Simulation results in Fig. 2(a) show the second neighbor distribution for nodes with $k = 4$ first neighbors in a network that follows power-law distribution with $N = 30\,\mathrm{K}$ and for $\alpha = 2.78$. The values of b_k and κ_k used were 0.0024 and 0.0013, respectively. The RMSE of the model values with the simulation results is 0.0052 and the corresponding NRMSE value is around 0.14. Thus the simulation result matches well with our derived model.

For the case Simulated-Gnutella network, we simulated the network with $N = 26870$ ultra-peer nodes. Figure 2(b) shows the second neighbor distribution of the nodes that have $k = 10$ first neighbors with a back edge probability, $b_k = 0.0043$ and cross edge probability $\kappa_k = 0.0002$. Here also the simulation result matches reasonably well with the model values, the RMSE being 0.0026 and the NRMSE is around 0.17.

5. Back Edge and Cross Edge Probabilities for Various Network Distributions

In this section, we initially propose an analytical model for deriving back/cross edge probabilities with respect to the degree of the nodes in uncorrelated random networks, with given arbitrary degree distribution. This kind of network can be formed by using the simple configuration model (see footnote b) where the degrees of nodes in the network are randomly chosen integers based on the specified distribution. Later, we derive the back/cross edge probabilities in random networks with given average clustering coefficient, c_k, of the peers of degree k, along with their

degree distribution. Such a topology can be generated randomly using the Serrano model described in Ref. 22.

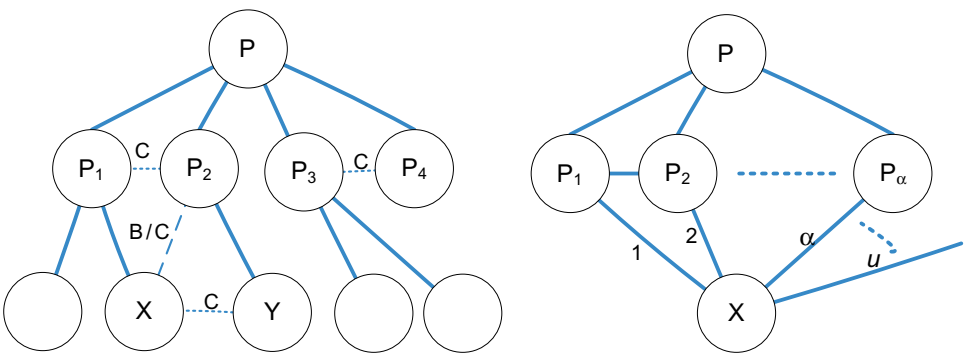
5.1. *Back and cross edge probabilities in uncorrelated random networks with arbitrary degree distribution*

We next derive the back and cross edge probabilities for uncorrelated random networks with given arbitrary distribution. As stated earlier, such a topology is considered to be randomly selected from the set of all topologies, with such a given distribution and can be generated using the configuration model discussed in footnote b. Prior to deriving the cross and back edge probabilities, we derive certain expressions that will be required for future use.

(1) The probability that any node in the network (say X) connects to a given node (say P_2) (see Fig. 3(b)) with degree v is given as $\frac{v}{N\langle z\rangle}$, where $\langle z\rangle$ represents the average degree of the network as found from Eq. (2). This is because, the probability that an edge from X leads to a random node of degree v is $\frac{vp_v}{\langle z\rangle}$. However since there are, on average, Np_v nodes of degree v, the probability that X connects to the given node, P_2, of degree v is

$$\left(\frac{vp_v}{\langle z\rangle}\right)\left(\frac{1}{Np_v}\right) = \frac{v}{N\langle z\rangle}.$$

(2) The probability that a random neighbor of P is of degree v is $\frac{vp_v}{\langle z\rangle}$; hence, the probability, ξ that a randomly selected edge from P_1 connects to any given



(a) A topology with cross/back edges (b) Cross/back edge formation

Fig. 3. (a) shows a portion of a p2p topology. The solid lines indicate the regular edges that connect two peers. The edges marked C are cross edges and the edge P_2X marked B/C is a cross edge for node X, where as it is a back edge for node P_2. (b) shows the formation of cross and back edges in networks with arbitrary degree distribution. A random node X connects to a random neighbor of P with probability $\xi = \frac{\langle z^2\rangle}{N\langle\langle z\rangle\rangle^2}$. Thus, a neighbor P_1 of P (with degree k) connects to any other neighbor of P with probability $(k-1)\xi$. The edges of a random node X are represented by a sequence number. Back edges are formed when a particular sequence of edges connects to more than one neighbors of P.

neighbor of P (see Fig. 3(b)) is

$$\xi = \sum_v \left(\frac{v}{N\langle z \rangle} \times \frac{vp_v}{\langle z \rangle} \right) = \frac{\langle z^2 \rangle}{N(\langle z \rangle)^2}, \tag{15}$$

where $\langle z^2 \rangle = \sum_v v^2 p_v$, represents the second moment of the node degree in the network.

We now initially derive the cross edge probability of a random node in the network, then go on to derive the back edge probability.

5.1.1. *Cross edge distribution in uncorrelated random networks with arbitrary degree distribution*

A random edge from a random neighbor of node P (node P is of degree k) is a cross edge, if the edge connects to any one of the other $k-1$ neighbors of P, the probability of which is $(k-1)\xi$. Thus the cross edge probability of a node with degree k, in any random network, is given as,

$$\kappa_k = (k-1)\xi = \frac{(k-1)\langle z^2 \rangle}{N(\langle z \rangle)^2}. \tag{16}$$

5.1.2. *Back edge probability in uncorrelated random networks with arbitrary degree distribution*

A random edge from P_1 (see Fig. 3(b)) is a back edge of node P with probability $\frac{\alpha-1}{\alpha}$, if the edge connects to a node X (level of X is greater than the level of P_1), such that X is connected to exactly $\alpha-1$ other neighbors of P. We derive the back edge probability by the sequence of following steps.

(1) We initially derive the probability that for a second neighbor, X, (of degree u) of node P (of degree k), a given sequence of $\alpha-1$ edges of X connects to any $\alpha-1$ neighbors of P (see Fig. 3(b)). The probability $\mathcal{P}(\alpha)$ of this is given as

$$\begin{aligned} \mathcal{P}(\alpha) &= (k-1)\xi(k-2)\xi \cdots (k-\alpha+1)\xi \\ &= (k-1)(k-2)\cdots(k-\alpha+1)\xi^{\alpha-1}. \end{aligned}$$

(2) For all possible sequence of edges of X, given that one edge is connected to a neighbor of P, we use the expression for $\mathcal{P}(\alpha)$ to next derive the probability, that out of the rest $u-1$ edges, exactly $\alpha-1$ edges connect to other neighbors of P. If $\mathscr{B}(\alpha)$ represents this probability, then

$$\mathscr{B}(\alpha) = \binom{k-1}{\alpha-1}\binom{u-1}{\alpha-1}\xi^{\alpha-1}(1-k\xi)^{u-\alpha}.$$

(3) Since α can vary from 1 to $\min(k,u)$, we next sum over all possible values of α; thus the probability that an edge from X to any neighbor of P is a back

edge is

$$\mathcal{B}_k(u) = \sum_{\alpha=1}^{\min(k,u)} \left[\frac{\alpha - 1}{\alpha} \mathscr{B}(\alpha) \right].$$

(4) The probability that the node X, reached through a neighbor of P is of degree u is $\frac{up_u}{\langle z \rangle}$; thus summing over all possible values of u, we get the probability that any random edge from a first neighbor of a k-degree node, P, is a back edge is,

$$b_k = (1 - \kappa_k) \sum_u \left[\frac{up_u}{\langle z \rangle} \mathcal{B}_k(u) \right]. \tag{17}$$

We next consider a special case of Erdős–Rényi Networks and derive their cross and back edge probabilities.

5.1.3. *Cross/back edge probability in Erdős–Rényi networks*

Equation 16 can be used to derive the cross edge probability in Erdős–Rényi networks; a special property of Erdős–Rényi network is that $\langle z^2 \rangle = (\langle z \rangle)^2$ (Ref. 18). Hence the cross edge probability of a random node of degree k in an Erdős–Rényi network is given by

$$\kappa_k = \frac{k-1}{N-2} \approx \frac{k-1}{N}. \tag{18}$$

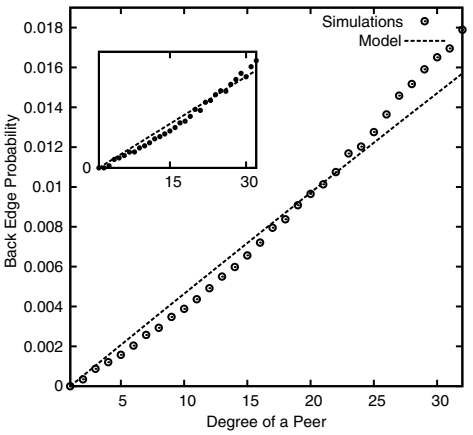
Similarly, Eq. (17) can be used to derive the back edge probability in Erdős–Rényi network; in this case, since $\langle z^2 \rangle = (\langle z \rangle)^2$, we replace ξ in Eq. (17) with $\frac{1}{N}$. Further, in these networks the probability that a random neighbor of node P_1 (here X as in Fig. 3(b)) is of degree u is simply p_{u-1} [17]; the back edge probability in Erdős–Rényi networks can be derived as

$$b_k = \sum_{\alpha=1}^{k} \binom{k-1}{\alpha-1} \hat{p}^{\alpha-1} \hat{q}^{k-\alpha} \cdot \left(1 - \frac{1}{\alpha}\right) = 1 - \frac{1}{k\hat{p}} \left(1 - \hat{q}^k\right), \tag{19}$$

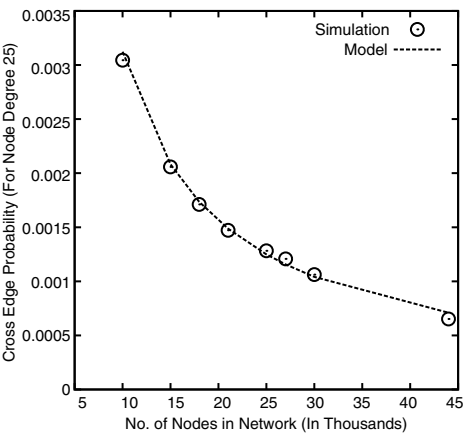
where $\hat{p}$ is the probability of connection of any two random nodes and $\hat{q} = 1 - \hat{p}$.

5.1.4. *Simulations*

We performed simulations to verify the correctness of our models using a random Gnutella network (shown in Fig. 4) and Erdős–Rényi networks (shown in Fig. 5(a)). The figure shows that theoretical output and simulation results are in good agreement. For the case of Erdős–Rényi networks, the RMSE values for cross and back edge probabilities are 0.0001 and 0.0004, respectively and the corresponding NRMSE values are 0.05 for both the cases. For the random Gnutella networks the RMSE and NRMSE values for the back edge probability are 0.00089 and 0.05, respectively, whereas for the cross edge probability the respective values are 0.000068 and 0.049. The results of both back and cross edges for random Gnutella network deviate a little at the higher degrees. Since the Gnutella network follows an heterogeneous degree distribution, the higher degrees cannot be made

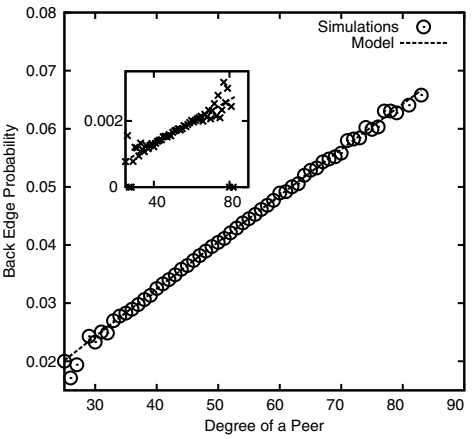


(a) Back/cross edge probability

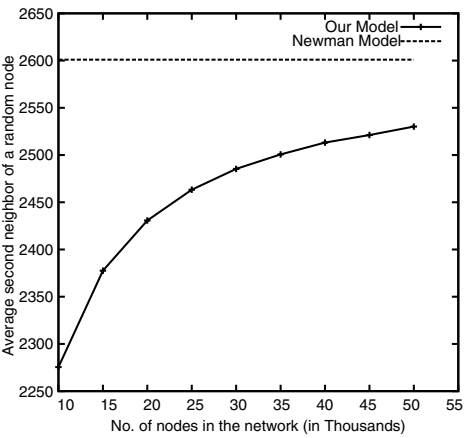


(b) Variation in cross edge probability with network size

Fig. 4. (a) The figure in the inset shows the back edge probability with respect to the degree of the nodes; the outside figure shows the cross edge probability for a network size of $N = 30\,\mathrm{K}$ nodes. The RMSE and NRMSE values for the back edge probability are 0.00089 and 0.05, respectively, whereas for the cross edge probability the respective values are 0.000068 and 0.049. (b) Shows the variation of cross edge probability for nodes with degree = 25 with varying network size.



(a) Back/cross edge probability



(b) Average no. of second neighbors

Fig. 5. (a) The figure in the inset shows the cross edge probability with respect to the degree of the nodes and the outside figure shows the back edge probability calculated using Eqs. (18) and (19), respectively. The RMSE are 0.0001 and 0.0004 for cross and back edge probabilities respectively and the corresponding NRMSE values are 0.05 for both the cases. (b) Compares the average second neighbor for the nodes calculated using our model (Eq. (14)) with that of the basic model (Eq. (6)). The network size is varied from 10 k to 50 k; however the average degree of the nodes are kept fixed and equal to 51.

completely uncorrelated through the configuration model, hence we observe the deviation in the result.

The model (Eqs. (16) and (17)) indicates that cross and back edge probabilities depend not only on the number of nodes in the network but also on the second moment of the nodes in the network. Hence, for a fixed second moment, the cross edge probabilities decrease with network size following a hyperbolic relation (Eq. (16)), as shown in Fig. 4(b). Consequently, as shown in Fig. 5(b), the average number of second neighbors of a node in the network, calculated using our model, is less than that calculated using the basic model, the difference is found to be around 5% for even a network as large as 40,000 nodes. That means $TTL(2)$ covers 5% less node than expected and produces similar amount of redundancy instead.

5.2. *Cross and back edge probabilities in clustered random networks*

Several researchers have argued that data sharing networks, including p2p networks like Gnutella and Kazaa, generally follow small world properties — showing random connectivity among the peers along with high clustering [14, 23, 24]. Hence, in this section we extend our model to derive cross and back edge probabilities in random networks with clustering [19]. We assume that the degree distribution and the first order clustering coefficient of the nodes of a particular degree is known, i.e. nodes with same degree are assumed to have the same clustering coefficient. We follow this assumption, as from the real Gnutella snapshots, we found that the clustering coefficients of the same degree nodes are almost same with very negligible variance. Existing p2p networks might reflect some higher order clustering [13] effects as well, i.e a tendency to form cycles of length greater than 3. However in our model we ignore such effects and assume that cycles greater than length 3 are formed due to random connectivity of the nodes. Thus apart from the degree distribution p_k of the network, let $\bar{c}(k)$ represent the average clustering coefficient of the nodes with degree k in the network. Such a network topology is assumed to be randomly selected from the set of all topologies with such a given degree distribution and clustering coefficients and can be generated using the Serrano model [22]. We next derive the cross and back edge probabilities respectively of these networks.

For a node P with degree k, the maximum number of triplets that can be formed is $\binom{k}{2}$. If $t(k)$ represents the average number of triangles formed by a node of degree k then according to the definition of clustering coefficient,

$$\bar{c}(k) = \frac{t(k)}{\binom{k}{2}} \Rightarrow t(k) = \binom{k}{2}\bar{c}(k).$$

For a node P, (see Fig. 3(a)) each triangle corresponds to a connectivity between two neighboring nodes of P (say P_1 and P_2). Thus the edge P_1P_2 can be traversed from P, either through P_1 or P_2, i.e. each such connecting edges can be traversed in 2 possible ways from the source node. If the average number of outgoing paths

from the neighbors of a k degree node (obtained by differentiating Eq. (4) at $x = 1$) is represented by $\langle s_k \rangle$, then the cross edge probability $\kappa_k^{(c)}$ of a node with degree k is given as

$$\kappa_k^{(c)} = \frac{2\bar{c}(k)}{\langle s_k \rangle} \binom{k}{2} = \frac{k(k-1)\bar{c}(k)}{\langle s_k \rangle}. \tag{20}$$

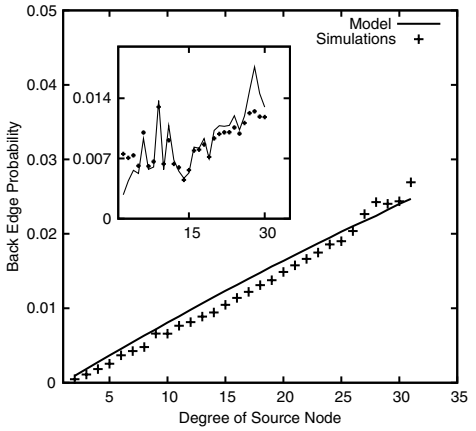
As we have assumed that the network connectivity is random with no higher order clustering, thus for a node P, apart from its cross edges, the other outgoing edges from the first neighbors of P connect to random nodes. Hence the back edge probability will be same as in the case of purely random networks and is given as in Eq. (17), with κ_k being replaced by $\kappa_k^{(c)}$. Thus the expression for the back edge probability for clustered random networks with any arbitrary degree distribution is given as

$$b_k^{(c)} = (1 - \kappa_k^{(c)}) \sum_u \left[\frac{up_u}{\langle z \rangle} \mathcal{B}_k(u) \right]. \tag{21}$$

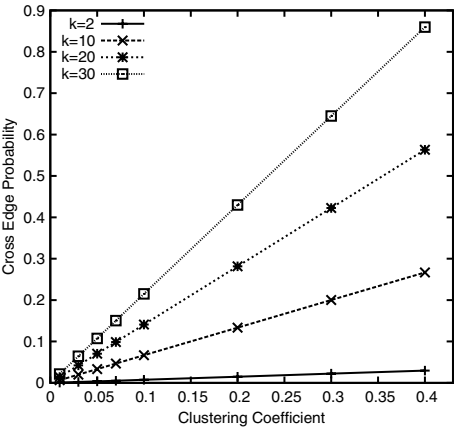
5.2.1. *Simulations*

We attempted to validate the models of back and cross edge probabilities in clustered random networks by considering the degree distribution and the average clustering coefficients of each degree class in real Gnutella topology (obtained from the real traces provided in [2]) and generating random networks with the same clustering properties using an algorithm proposed by Serrano *et al.* [22]. We refer this topology as the *Serrano–Gnutella* topology. The number of nodes considered for the simulation was around 23 K and the average clustering coefficient of the whole network was 0.03. The results shown in Fig. 6(a) indicates that the model provides a good estimate of the back and cross edge probabilities in random clustered networks, where RMSE of both the model values of back and cross edge probabilities with respect to the simulation result is 0.001 and the NRMSE values are 0.05 and 0.14, respectively. We show the effect of increasing clustering coefficient on the cross edge probability of the peers of various degrees. As can be seen in Fig. 6(b), the cross edge probability of the high degree peers increases drastically with increasing clustering coefficient, as compared to the low degree peers. This indicates that increasing the clustering coefficient has a huge impact on the coverage of the high degree peers. Further, we also simulated the average $TTL(2)$ coverage of the peers in a network for a given node degree; as shown in Fig. 6(c), our model provides a good estimate of the average $TTL(2)$ coverage of a random peer of a given degree. In this case, the RMSE value of the model values with respect to the simulation results is around 17.09 and the NRMSE value is 0.041 that indicates that the model values matches well with the simulation results.

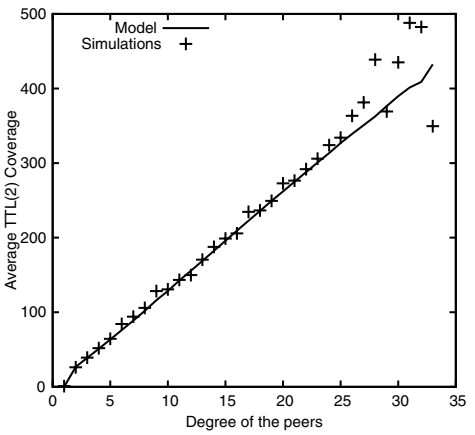
In the next section, we derive the network coverage of the peers for any generic TTL value and show how the back/cross edge probabilities of a node increase drastically with increasing distance from the source node.



(a) Back/cross edge probability



(b) Effect of clustering coefficient



(c) Average no. of second neighbors

Fig. 6. (a) shows the back edge probabilities of the nodes, for various peer degrees, in the random clustered network (Serrano–Gnutella) with 23 K nodes and an average clustering coefficient of 0.03, for both simulations and the model results. The figure in the inset compares the cross edge probabilities. RMSE for the model values of both back and cross edge probabilities with respect to the simulation results is 0.001, where as the NRMSE values are 0.05 and 0.14, respectively. (b) shows the variation of cross edge probability with average clustering coefficient for various node degrees. (c) shows the $TTL(2)$ coverage of the peers (i.e. the average number of second neighbors) with respect to the degree of the peers in the same network (RMSE = 17.09).

6. Generalizing the Back/Cross Edge Probabilities for Higher TTL Values

In this section we derive the expressions for the level l back and cross edge probabilities of the peers. Let $\kappa_k(l)$ and $b_k(l)$ denote the cross and back edge probability at level l of a node of degree k. The values of $\kappa_k(l)$ and $b_k(l)$ determine the average

number of $(l+1)$th hop neighbors of a node. If $a_k(l)$ denotes the average network coverage of a source node, of degree k, up to level l from it, then $a_k(l) - a_k(l-1)$ is the number of nodes at level l only of the node. Although it is difficult to find a closed form expression of $a_k(l)$, however, we can derive a recursive equation of it. The number of unique nodes reached through the edges at level l, of a degree k source node, is $\langle z\rangle(a_k(l) - a_k(l-1))(1 - b_k(l) - \kappa_k(l))$, and hence

$$a_k(l+1) = a_k(l) + \langle z\rangle(a_k(l) - a_k(l-1))(1 - b_k(l) - \kappa_k(l)), \tag{22}$$

and the average number of level l neighbors of any random node is given as $\hat{N}(l) = \sum_k (a_k(l) - a_k(l-1))p_k$. We next derive expressions for $b_k(l)$ and $\kappa_k(l)$, at a level l from the source node for the case of networks with random peer connectivities.

6.1. $\kappa_k(l)$ *in networks with random peer connections*

Analogous to the derivation of the level 1 cross edge probability in Eq. (16), the cross edge probability at level l, of a k-degree node can be represented as

$$\kappa_k(l) = \frac{(a_k(l) - a_k(l-2))\langle z^2\rangle}{(N - a_k(l-2))(\langle z\rangle)^2} \tag{23}$$

where $a_k(l) - a_k(l-2)$ gives the average number of nodes at level l and $l-1$, of a k-degree source node, and $N - a_k(l-2)$ is the possible number of nodes to which an edge from a level l node can connect.

6.2. $b_k(l)$ *in networks with random peer connections*

Analogous to the derivation of back edge probability in Eq. (17), the back edge probability at level l of a source node of degree k, is given as,

$$b_k(l) = (1 - \kappa_k(l)) \sum_u \left[\frac{up_u}{\langle z\rangle} \mathcal{B}_{a_k(l) - a_k(l-1)}(u) \right] \tag{24}$$

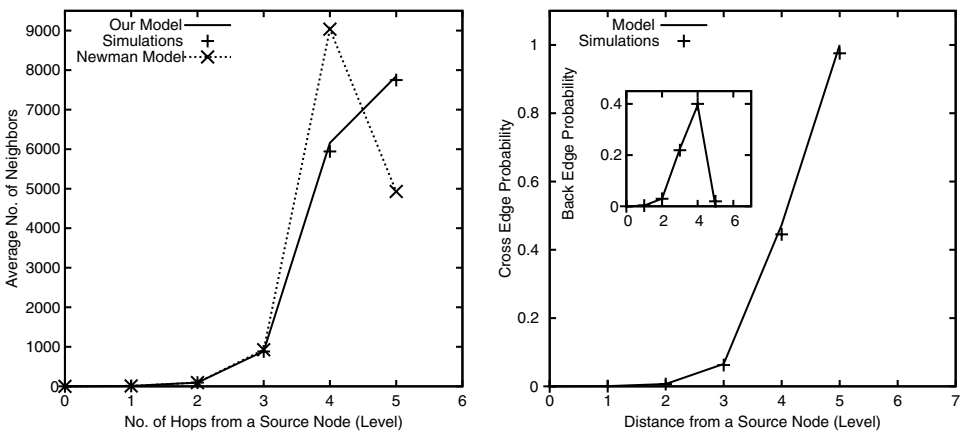
where, $\mathcal{B}_{a_k(l)-a_k(l-1)}(u)$ is calculated using Eq. (17) by replacing k with $a_k(l) - a_k(l-1)$. The base conditions for Eq. (22) are the average network coverage of a random peer up to level 0, 1 and 2, respectively and are as follows:

$$a_k(0) = 1, \quad a_k(1) = k + 1, \quad a_2 = \hat{S}'_k(1) + \langle z\rangle + 1, \tag{25}$$

where $\hat{S}'_k(1)$ is the average number of second neighbors of a source node, with degree k, in the presence of back/cross edges derived from Eq. (13). The back and cross edge probabilities at level 0 are same and equal to 0; the values at level 1 can be computed from Eqs. (17) and (16), respectively. Although it is difficult to find an exact solution of Eq. (22), however, we can iteratively calculate the values of $a_k(l)$. We next present the results of the validation of the model using simulations.

6.3. *Simulation results and discussion*

Figure 7(a) shows the average number of nodes at each level for a random source node in an Erdős–Rényi network of 15 K nodes with an average degree of 9.75. The average values of the neighbors at each level, calculated using our model, are validated using the simulation results. For comparisons, we also present the average number of neighbors at various levels according to the basic Newman model. Figure 7(a) shows that the average number of neighbors at various levels predicted using our model matches well with the simulation results as compared to the Newman model. The RMSE values of our model and the Newman model as compared to the simulation results are 96.85 and 1708.82, respectively while the corresponding NRMSE values are 0.012 and 0.218. The difference in coverage between basic and our model seems to increase exponentially as level increases. From 5–10% difference observed in level 2, the difference becomes as high as 50% at level 4. (The drop is seen at level 5 as there is no remaining nodes to be explored in this particular example, i.e. the diameter of the network is 5). This difference can be understood by looking at the average back and cross edge probabilities at various levels (which is in perfect agreement). Figure 7(b) shows them as derived from the model and simulations. The RMSE as well as the NRMSE values of the cross edge probabilities



(a) Back/cross edge probabilities at every levels

(b) Number of neighbors at every level

Fig. 7. (b) The figure in the inset shows the back edge probability with respect to the degree of the nodes, the outside figure shows the cross edge probability. The RMSE as well as the NRMSE value of the cross edge probabilities of our model with respect to the simulation results is 0.015, whereas the RMSE and the NRMSE values of the back edge probabilities are 0.008 and 0.021 respectively. (a) shows the average number of neighbors of a randomly selected node. The size of the network is 15k and the connection probability $\hat{p} = 0.00065$. The RMSE values of our model and the Newman model as compared to the simulation results are 96.85 and 1708.82, respectively while the corresponding NRMSE values are 0.012 and 0.218.

of our model with respect to the simulation results are same and equal to 0.015, whereas the RMSE and the NRMSE values of the back edge probabilities are 0.008 and 0.021, respectively, that indicates a very good fit of the model values with the simulation results. It can be observed, beyond a particular threshold level (3, in this case), the back and cross edge probabilities increase enormously at a very high rate.

Since the short cycles induced by the back and cross edges generates duplicate messages, it is observed that the redundancy in flooding increases enormously with TTL value due to very high cross/back edge probabilities at higher levels. However, for smaller TTL, the coverage may not be high and may be unsuitable for various activities. For example, the percentage of nodes covered up to level 3 (by using a $TTL(3)$ message) is merely around 6.5% of the total network size while the redundancy at level 4 is around 50% that is, every three message packets can only find two unexplored nodes. Thus the model provides a tool to the designers to explore the tradeoff between network coverage and traffic redundancy for a given network and decide accordingly.

7. Conclusion

In this paper, we have developed suitable models that quantify the coverage of the peers in networks performing TTL based searches. The models based on generating function formalism provide a strong theoretical foundation needed to understand the relation between the topology of a network and the achievable performance through TTL-based searches. Using the derived models, we have provided an insight of the effects of back and cross edges on the network coverage of the peers. Through this formalism each individual peer can easily estimate global properties like back edge, cross edge, and hence network coverage, from the degree distribution of the network (which can be predicted in many cases or can be estimated using samples collected from nearby nodes) and local properties like its own degree and clustering coefficient. Although the models have been derived for p2p networks, however, flooding is a generalized phenomenon and finds wide use in social and information networks. Thus these bounds can be suitably applied for these networks also. However, the derived models are based on a basic assumption that there is no correlation of the connectivity between two nodes and other parameters like node degree, node strength etc. But in many practical networks (like social networks), these correlations exist and also play an important role in determining the network coverage. Further, in our derivations, we have also ignored the existence of some motifs, like existence of quadrilaterals, in the network that can also play an important role in determining the coverage of the peers in the network. Thus, an important future direction of research would be to study these network parameters for correlated (social) networks and develop suitable strategies to improve network coverage besides reducing traffic redundancy.

References

[1] How gnutella works? http://wiki.limewire.org.

[2] Multimedia and Internetworking Research Group, University of Oregon, http://mirage.cs.uoregon.edu.

[3] Aggarwal, V., Feldmann, A. and Scheideler, C., Can ISP's and P2P users cooperate for improved performance? *SIGCOMM Comput. Commun. Rev.* **37** (2007) 29–40.

[4] Albert, R. and Barabasi, A.-L., Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74** (2002) 47.

[5] Barabasi, A. L. and Albert, R., Emergence of scaling in random networks, *Science* **286** (1999) 509–512.

[6] Bender, E. A. and Canfield, E. R., The asymptotic number of labeled graphs with given degree sequences, *J. Comb. Theory, Ser. A* **24** (1978) 296–307.

[7] Chandra, J., Shaw, S. and Ganguly, N., Analyzing network coverage in unstructured peer-to-peer networks: A complex network approach, in *Proceedings of Networking 2009* (2009), pp. 690–702.

[8] Chandra, J., Shaw, S. K. and Ganguly, N., HPC5: An efficient topology generation mechanism for gnutella networks, *Comput. Netw.* **54** (2010) 1440–1459.

[9] Day, K. and Tripathi, A., A comparative study of topological properties of hypercubes and star graphs, *IEEE Transactions on Parallel and Distributed Systems* **5** (1994) 31–38.

[10] Donato, D., Laura, L., Leonardi, S. and Millozzi, S. Large scale properties of the webgraph, *Eur. Phys. J. B* **38** (2004) 239–243.

[11] Faloutsos, M., Faloutsos, P. and Faloutsos, C., On power-law relationships of the internet topology, *SIGCOMM Comput. Commun. Rev.* **29** (1999) 251–262.

[12] Fletcher, G., Sheth, H. and Brner, K., Unstructured peer-to-peer networks: Topological properties and search performance, in *Agents and Peer-to-Peer Computing*, Lecture Notes in Computer Science, Vol. 3601 (Springer Berlin/Heidelberg, 2005), pp. 14–27.

[13] Fronczak, A., Holyst, J. A., Jedynak, M. and Sienkiewicz, J., Higher order clustering coefficients in Barabasi-Albert networks, *Physica A* **316** (2002) 688–694.

[14] Iamnitchi, A., Ripeanu, M. and Foster, I. T., Small-world file-sharing communities, in *INFOCOM: The Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies* (2004).

[15] Karbhari, P., Ammar, M. H., Dhamdhere, A., Raj, H., Riley, G. F. and Zegura, E. W., Bootstrapping in gnutella: A measurement study, in *PAM*, *Lecture Notes in Computer Science*, Vol. 3015 (Springer, 2004), ISBN 3-540-21492-5, pp. 22–32.

[16] Mitra, B., Dubey, A. K., Ghose, S. and Ganguly, N., How do superpeer networks emerge? in *Proceedings of the 29th Conference on Information Communications*, INFOCOM'10 (IEEE Press, Piscataway, NJ, USA, 2010), ISBN 978-1-4244-5836-3, pp. 1514–1522, http://portal.acm.org/citation.cfm?id=1833515.1833730.

[17] Newman, M. E., Strogatz, S. H. and Watts, D. J., Random graphs with arbitrary degree distributions and their applications, *Phys. Rev. E* **64** (2001).

[18] Newman, M. E. J., The structure and function of complex networks, *SIREV: SIAM Review* **45** (2003).

[19] Newman, M. E. J., Random graphs with clustering, *Phys. Rev. Lett.* **103** (2009) 058701.

[20] Ripeanu, M. and Foster, I. T., Mapping the gnutella network: Macroscopic properties of large-scale peer-to-peer systems, in *Revised Papers from the First International Workshop on Peer-to-Peer Systems*, IPTPS '01 (Springer-Verlag,

London, UK, 2002), ISBN 3-540-44179-4, pp. 85–93, http://portal.acm.org/citation.cfm?id=646334.687818.

[21] Sen, S. and Wang, J., Analyzing peer-to-peer traffic across large networks, in *IMW '02: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurement* (ACM, New York, NY, USA, 2002), pp. 137–150.

[22] Serrano, M. and Boguná, M., Tuning clustering in random networks with arbitrary degree distributions, *Phys. Rev. E* **72** (2005).

[23] Stutzbach, D. and Rejaie, R., Capturing accurate snapshots of the gnutella network, in *INFOCOM: The Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies* (2005), pp. 2825–2830.

[24] Stutzbach, D., Rejaie, R. and Sen, S., Characterizing unstructured overlay topologies in modern P2P file-sharing systems, in *Internet Measurment Conference* (USENIX Association, 2005), pp. 49–62.

[25] Stutzbach, D., Rejaie, R. and Sen, S., Characterizing unstructured overlay topologies in modern p2p file-sharing systems, *IEEE/ACM Transactions on Networking* **16** (2008) 267–280.

[26] Vázquez, A., Pastor-Satorras, R. and Vespignani, A., Large-scale topological and dynamical properties of the internet, *Phys. Rev. E* **65** (2002) 066130.

[27] William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd edn. (John Wiley and Sons, 2000).

[28] Zhenzhou, Z., Panos, K. and Spiridon, B., DCMP: A distributed cycle minimization protocol for peer-to-peer networks, in *Parallel and Distributed Systems, IEEE Transactions* **19** (2008) 363–377.