Analysis models for unguided search in unstructured P2P networks

Bin Wu and Ajay D. Kshemkalyani*

Computer Science Department,
University of Illinois at Chicago,
Chicago, IL 60607, USA
E-mail: bwu@cs.uic.edu
E-mail: ajayk@cs.uic.edu
*Corresponding author

Abstract: Random walk and flooding are basic mechanisms for searching unstructured overlays. This paper shows that node coverage is an important metric for query performance in random graph based P2P networks. We then present two analytical models: the algebraic model and the combinatorial model. These models are useful in setting query parameters and evaluating search efficiency. We evaluate these two models in terms of various performance metrics against simulation results. We also study the impact of the setup parameters, such as node degree, hop count, and number of walkers, on the query performance, and on the precision of the models.

Keywords: analysis models; unguided search; unstructured P2P.

Reference to this paper should be made as follows: Wu, B. and Kshemkalyani, A.D. (2008) 'Analysis models for unguided search in unstructured P2P networks', *Int. J. Ad Hoc and Ubiquitous Computing*, Vol. 3, No. 4, pp.255–263.

Biographical notes: Bin Wu received his BS and MS Degree from the Department of Energy Engineering at Zhejiang University, China, in 1992 and 1995, respectively. He received the MS Degree and PhD Degree in Computer Science from the University of Illinois at Chicago in 2002 and 2006, respectively. His research interests include distributed computing, P2P networks, network protocols and algorithms, and web-based healthcare information technologies.

Ajay D. Kshemkalyani received PhD Degree in Computer and Information Science from Ohio State University in 1991 and the BTech Degree in Computer Science and Engineering from the Indian Institute of Technology, Bombay, in 1987. His research interests are in computer networks, distributed computing, algorithms and concurrent systems. He has been an Associate Professor at the University of Illinois at Chicago since 2000, before which he spent several years at IBM Research Triangle Park working on various aspects of computer networks. He is a member of the ACM and a senior member of the IEEE. In 1999, he received the US National Science Foundation's CAREER Award. He is currently on the Editorial Board of the *Elsevier Journal, Computer Networks*.

1 Introduction

Peer-to-peer networks aim to allow internet users to loosely organise themselves to share their resources with ease of implementation and maintenance (Eng et al., 2005). Unstructured peer-to-peer overlays have the following advantages over structured overlays (Risson and Moors, 2006): they can handle high churn rates easier, they do not incur much overhead for maintaining the logical structure, and they support keyword searches (Liu et al., 2004; Reynolds and Vahdat, 2003) based on semantic identification and information retrieval techniques (Tang et al., 2003), and complex queries such as range queries. More realistic P2P applications have recently been developed with unstructured overlays than with structured ones due to these advantages.

Recent research on unstructured P2P networks has focused on the search strategies and replication schemes (Eng et al., 2005; Lv et al., 2002; Risson and Moors, 2006). The goal is to avoid message explosion, achieve good performance, and maintain the simple and flexible topology of an unstructured P2P overlay. The improvements over the traditional flooding are the expanding ring flooding and random walk (Lv et al., 2002). These approaches can be classified as *unguided searches*. In contrast, *guided searches* remember some specific information on the network topology or on the past searches. When forwarding a query message in search of an object, this information is used to narrow down the choice of the neighbour(s) to forward the query to Crespo and Garcia-Molina (2002) and Tsoumakos and Roussopoulos (2003).

In unstructured overlays, an object is not usually identified by its unique ID, and it may not even have an object ID. Thus, query by keyword is a primary method of indexing and searching for objects in such environments (Liu et al., 2004; Reynolds and Vahdat, 2003). For keyword searches, the 'matching' depends on the relevancy between an object and the set of keywords used in the query. Keyword based search methods are usually closely related to semantic identification and information retrieval techniques (Schmitz, 2004; Tang et al., 2003; Zeinalipour-Yazti et al., 2004) such that an effective semantic based clustering and a pertinent model on the characteristics and distributions of keywords (Bawa et al., 2003) are essential.

Existing studies of blind search are based on empirical rationale and the performance of blind search is studied primarily through simulations. Using extensive modelling and simulations, Lv et al. (2002) studied the performance of P2P systems under three considerations: network topology, query distribution, and replication distribution. Based on models from queuing theory, (Ge et al., 2003) presented the performance of different indexing approaches, in terms of system throughput and probability of 'successful query'. A study of random walk based on graph theory and Chernoff bounds compared the performance of random walk with that of flooding (Gkantsidis et al., 2006). This study by Gkantsidis et al. shows that the effect of a k-step random walk is statistically similar to that of taking k independent samples in a well-connected graph. Using this approximation of independent sampling, simplified formulae for the *success rate*, *message overhead*, and *time overhead* of random walk as functions of Time to Live (TTL), object popularity, and number of walkers were given (Bisnik and Abouzeid, 2005). The *success rate* was statistically characterised by the probability of a successful search, $p_s = 1 - (1 - c)^{kT}$, where c is the popularity of the object (i.e., the ratio of the nodes that have an object copy), T is the TTL, and k is the number of random walkers. This approach relies on an accurate estimate of the popularity, which is usually not available.

Contributions

- This study shows that *node coverage* is a useful metric in analysing the performance of blind searches – specifically, to estimate the message efficiency, success rate, and object recall. *Node coverage* is defined as the fraction of nodes visited in a query search. There has been no analytic model for node coverage computation.
- We then formulate two simple theoretical models for analysing the characteristics of search methods in unstructured P2P networks. These models are based on node coverage analysis on the *random graph* topology, which is a small-world network. The models provide a guideline to understand how the settings of the querying parameters and network characteristics impact the efficiency of the search strategies. This will allow system designers to tune parameters to achieve desired performance trade-offs and to estimate object replication on a statistical basis.
- We evaluate the goodness of two analytical models – the algebraic model and the combinatorial model – against simulation results, for various search efficiency metrics. We use the random graph topology and assume unguided searches. The search metrics we consider are the node coverage, the message efficiency, and the object recall. The results show that the two analytical models are very close to each other and reasonably accurate. Using simulations, we evaluate the impact of the parameters such as the average node degree, hop count, number of walkers, and replication ratios on node coverage, object recall, and message efficiency, and on the accuracy of the models. The preliminary results were presented in Wu and Kshemkalyani (2006a, 2006b).
- A simple method of estimating search success rate was proposed in Gkantsidis et al. (2006). We compare this method for computing success rate with the success rate computed from our node coverage models. This will help to discover more factors that affect the effectiveness of our analysis models.

2 Node coverage

In a random graph overlay where object distributions are also random, the more nodes that a query process covers, the higher the chance that a specified object is found, and the more the expected number of copies of the expected objects that can be retrieved. We define *node coverage* as the fraction of nodes visited by query messages. We express it as a function of the message overhead. Node coverage is independent of object characteristics and depends only on the search strategy and the graph topology. We identify the following uses of node coverage.

- Node coverage gives a more reasonable basis to calculate *success rate*, defined as the probability that a search finds a satisfactory object within the specified constraints, such as the number of hops or the message overhead. Knowing such answers is useful in setting the querying parameters such as TTL or the number of walkers for the search.
- For keyword searches and range searches, it is often desirable to find as many objects as possible that satisfy the search criteria. The number of qualified objects that are found is defined as *object recall*. *Message efficiency*, defined as the number of qualifying objects retrieved (i.e., recall) per query message, is another important performance metric in evaluating querying methods. Node coverage can also serve as an indicator of object recall.

- It may happen that a queried object does not exist in the network. A high node coverage without a query success indicates a high likelihood of this condition. This can be used as a guideline to call off the search.

The notation used in this paper is given in Table 1.

Table 1 Notations for the models

N	Total number of nodes
p	Probability of a link between two nodes (random graph)
d	Average degree of a node
w	No. of random walkers
r	No. of object replicas
x	No. of query messages
h	No. of query hops
v	No. of visits to nodes
u, $u(x)$, $u(h)$, $u(h, w)$	No. of distinct nodes visited
u/N	Node coverage
$p_s(h, w, r)$	Success probability for a search

3 The algebraic model

This model performs a node coverage analysis but makes no distinction among the search methods when node coverage is computed. Each query message is treated as an independent sample. This model gives the expected node coverage in terms of the message overhead, and then in terms of the hop count.

The first hop of message forwarding always covers $w+1$ distinct nodes in random walk, or $d+1$ expected nodes in flooding (including the initiator). Due to the randomness of node links, from the second hop onwards, a message forwarding may visit a node that has already been visited.

Suppose a randomly chosen link is being probed by a message, and u is the expected number of distinct nodes that have been visited so far. The probability that a new node is discovered by this message is $\frac{N-u}{N-2}$. Thus, the expected number of distinct nodes visited would be $u+\frac{N-u}{N-2}$ after this message.

Let x denote the number of query messages so far. Then:

$$u(x+1) = u(x) + \frac{N-u(x)}{N-1} \tag{1}$$

which can be approximated as:

$$u'(x) = \frac{N}{N-2} - \frac{u(x)}{N-2}. \tag{2}$$

This equation can be solved as:

$$u(x) = Ce^{\frac{-x}{N-2}} + N. \tag{3}$$

Here C is a constant determined by the initial condition.

Random Walkers. Within the first hop, $w+1$ distinct nodes (including the initiator itself) are visited. The initial condition takes the following form:

$$u(w+1) = w+1. \tag{4}$$

We can then solve for the constant C:

$$C = (w+1-N)e^{\frac{w+1}{N-2}}. \tag{5}$$

Then the complete solution for equation (3) is:

$$u(x) = \begin{cases} N-(N-w-1)e^{\frac{w+1-x}{N-2}} & \text{if } x > w+1 \\ x & \text{if } x \le w+1 \end{cases} \tag{6}$$

If we express u in terms of h, where h is the number of query hops and $x = wh+1$, we obtain:

$$u(x) = \begin{cases} N-(N-w-1)e^{\frac{w(1-h)}{N-2}} & \text{if } h > 1 \\ w+1 & \text{if } h = 1 \end{cases} \tag{7}$$

Flooding. The difference between $u(x)$ for random walk and for flooding is that for flooding, message overhead is exponential in number of hops.

The initial condition for equation (3) for flooding is:

$$u(d+1) = d+1. \tag{8}$$

We can then solve for the constant C:

$$C = (d+1-N)e^{\frac{d+1}{N-2}}. \tag{9}$$

Then the complete solution for equation (3) is:

$$u(x) = \begin{cases} N-(N-d-1)e^{\frac{d+1-x}{N-2}} & \text{if } x > w+1 \\ x & \text{if } x \le w+1 \end{cases} \tag{10}$$

As $x = \sum_{i=0}^{h} d^i = \frac{d^{h+1}-1}{d-1}$, we can express u in terms of h as:

$$u(h) = \begin{cases} N-(N-d-1)e^{\frac{d^2-d^{h+1}}{(d-1)(N-1)}} & \text{if } h > 1 \\ d+1 & \text{if } h = 1 \end{cases} \tag{11}$$

Note that this model makes little distinction between random walk and flooding. They are just special cases of equation (3) with different initial conditions and different expressions of h in terms of x. The algebraic model is simple and expected to have relatively low precision.

Replication. We assume random replication – r copies of an object are randomly distributed in the network. The probability of finding a copy (success rate) then becomes:

$$p_s(h) = 1 - \left(1 - \frac{u(h)}{N}\right)^r. \tag{12}$$

The term $u(h) = N$ is the node coverage.

4 Combinatorial model for random walk

In this model, we directly estimate the *node coverage* in terms of the number of message hops. The coverage analysis begins by analysing the behaviour of a single random walker and then extends the results to multiple walkers. The behaviour of a single walker can be treated as a random sampling and multiple walkers are considered to be independent. For a single random walker, if we consider the node coverage as a state variable, the state after the next hop depends only on the current state. The walk can thus be modelled as a Markov process.

Let v be the number of nodes visited so far. Let $Pr(u, v)$ denote the probability that after v node visits, u distinct nodes have been visited.

For the first hop $(v = 2)$:

- $Pr(2, 2) = 1$.

For the second hop $(v = 3)$:

- $Pr(2, 3) = Pr(2, 2) \cdot 0 = 0$
- $Pr(3, 3) = Pr(2, 2) \cdot 1 = 1$.

For the third hop $(v = 4)$:

- $Pr(3,4) = Pr(3,3) \cdot \frac{3-2}{N-2}$
- $Pr(4,4) = Pr(3,3) \cdot \frac{N-3}{N-2}$.

For the fourth hop $(v = 5)$:

- $Pr(3,5) = Pr(3,4) \cdot \frac{3-2}{N-2}$
- $Pr(4,5) = Pr(4,4) \cdot \frac{4-2}{N-2} + Pr(3,4) \cdot \frac{N-3}{N-2}$
- $Pr(5,5) = Pr(4,4) \cdot \frac{N-4}{N-2}$.

For the fifth hop $(v = 6)$:

- $Pr(3,6) = Pr(3,4) \cdot \frac{3-2}{N-2}$
- $Pr(4,6) = Pr(4,5) \cdot \frac{4-2}{N-2} + Pr(3,5) \cdot \frac{N-3}{N-2}$
- $Pr(5,6) = Pr(4,5) \cdot \frac{5-2}{N-2} + Pr(4,5) \cdot \frac{N-4}{N-2}$
- $Pr(6,6) = Pr(5,5) \cdot \frac{N-5}{N-2}$.

Based on this pattern, the inductive expression for probability $Pr(u, v)$ is given in Figure 1.

Figure 1 $Pr(u, v)$ for the combinatorial model

$$Pr(u,v) \text{ (for } u \le v) = \begin{cases} 1 & \text{if } u = 2, v = 2 \\ Pr(u, v-1) \cdot \frac{u-2}{N-2} & \text{if } u = 2, v \neq 2 \\ Pr(u-1, v-1) \cdot \frac{N-(u-1)}{N-2} + Pr(u, v-1) \cdot \frac{u-2}{N-2} & \text{if } 2 < u < v \\ Pr(u-1, v-1) \cdot \frac{N-(u-1)}{N-2} & \text{if } u = v > 2 \end{cases} \tag{13}$$

Let $\overline{u}(h, w)$ denote the expected number of distinct nodes covered by w random walkers after travelling h hops ($h + 1$ messages). Then $\overline{u}(h,1)$ can be expressed as follows,

$$\overline{u}(h,1) = \begin{cases} h+1 & \text{if } h \le 2 \\ \sum_{i=3}^{h+1} Pr(i, h+1) \cdot i & \text{if } h > 2 \end{cases} \tag{14}$$

which is the average of all possible number of distinct nodes, weighted by their respective probabilities, given the number of hops h.

Assuming there are r copies of the desired object, the success rate for a single walker is expected to be:

$$p_s(h,1,r) = 1 - \left[1 - \frac{\overline{u}(h,1)}{N}\right]^r. \tag{15}$$

As multiple $(w > 1)$ walkers travel independent of each other, any walker is expected to visit $\overline{u}(h,1)$ 'distinct' nodes after h hops. Here, 'distinct' refers to the nodes witnessed by a single walker. It is possible that some of those $\overline{u}(h,1)$ nodes have been visited by other walkers. To compute $\overline{u}(h, w)$ for $w > 1$, assume they travel the network sequentially.

- After the first walker finishes, $\overline{u}(h,1)$ distinct nodes were visited. Let $new(h,1) = \overline{u}(h,1)$.
- The second walker sees $\overline{u}(h,1)$ distinct nodes according to its own witness. Among those $\overline{u}(h,1)$ nodes, $\left(1 - \frac{\overline{u}(h,1)}{N}\right)\overline{u}(h,1)$ are expected to be new from the previous walk, which is denoted as $new(h, 2)$.
- For the ith walker, the expected number of new nodes it discovers is:

$$new(h,i) = \left[1 - \frac{\sum_{k=1}^{i-1} new(h,k)}{N}\right]\overline{u}(h,1). \tag{16}$$

- The expected total number of nodes that are visit by w random walkers, after h hops, is:

$$u(h,w) = \sum_{i=1}^{w} new(h,i). \tag{17}$$

- The success rate can be expressed using node coverage:

$$p_s(h,w,r) = 1 - \left(1 - \frac{\overline{u}(h,w)}{N}\right)^r. \tag{18}$$

This model has computational complexity $O(N^2)$ due to the nature of equations (13) and (14). In contrast, the algebraic model has complexity $O(1)$.

5 Evaluation of the algebraic and combinatorial mode

To test the validity of the algebraic and combinatorial models, we simulated random walk on a undirected random graph having $N = 20{,}000$ nodes. We measured the node coverage, object recall, and message efficiency under different settings of querying parameters and node degree. Both models performed very close to each other; hence in the graphs for most experiments, we show their values as a single plot. A study of the small differences between the models is shown in Section 5.4.

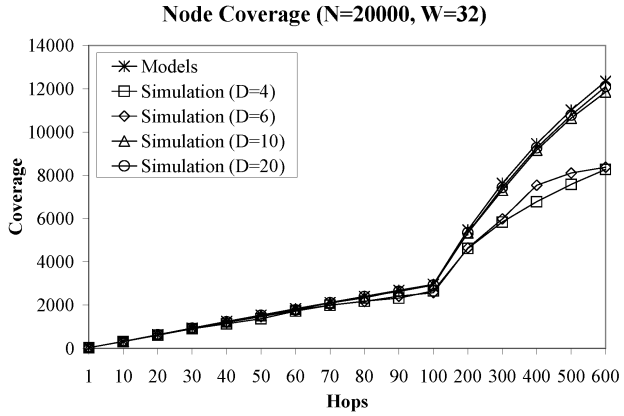
5.1 Evaluation on node coverage

We compare node coverage as computed from our analytic models and that obtained from the simulations. We study the impact of d and w on node coverage.

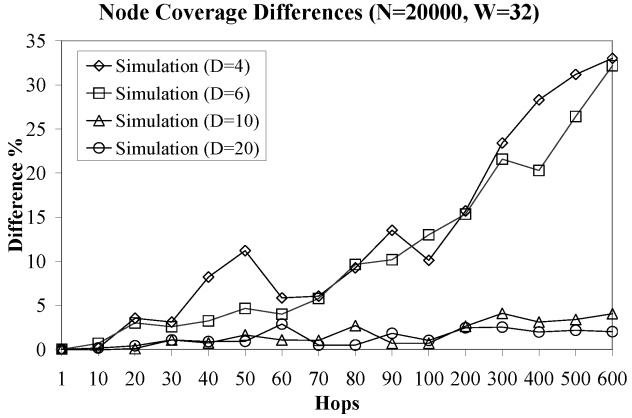
Effect of node degree. We simulated 32 random walkers and varied the average node degree. From Figure 2(a), observe that both models give higher values of node coverage than the simulation results. When the node degree is small ($d = 4$, $d = 6$) the difference between analytical results and simulations is large (up to over 30% for $h = 600$) but this difference reduces as d increases. For $d = 20$, the difference remains below 3% at all values of h (see Figure 2(b)).

The analytical models consider each 'next step' of message forwarding as a random sampling and the probability of visiting a new node is only determined by the current node coverage; the node degree is not taken into account. Consider a node i that is now being visited for the second time, when it forwards this message, ideally it should select a neighbour that was not visited before. However, as our models assume a stateless search, i may forward the message along an already explored link. This reduces the probability of forwarding to an unvisited node. Since the analytical models do not account for this effect, the coverage computed from our models is always higher than simulation results. However, this effect diminishes as the average node degree increases because for a higher degree node, the chance that a subsequent forwarding goes through a fresh link is higher than that for a low degree node.

Figure 2 Impact of average node degree on node coverage: (a) absolute values and (b) relative differences with respect to models



(a)

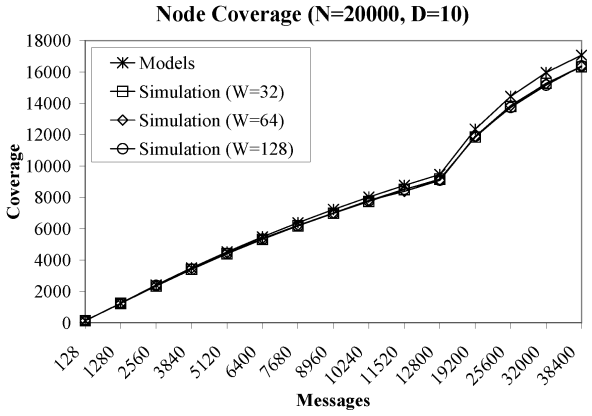


(b)

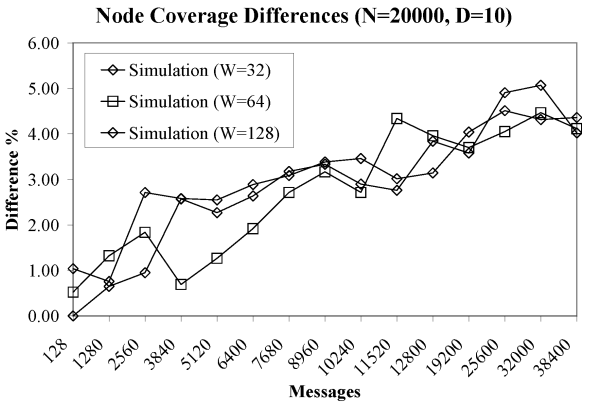
Observe from Figure 2(b) that for all values of node degree, the differences between analytical and simulation results increase as the hop count h increases. This is because as the hop number increases, the possibility that the node selects a 'dirty link' also increases, magnifying the effect of node degree on the outcoming node coverage. Note that the analytical models assumes that the node always selects a 'fresh' link to forward a message.

Effect of the number of walkers. Figure 3(a) shows the comparison of node coverage for 32, 64, and 128 walkers in the random graph with $d = 10$. As observed above, the simulation results are consistently smaller than our analytical results and the difference increases as the message number increases. Our simulations show that the number of walkers has no significant impact on node coverage. Figure 3(b) shows the relative deviation of each of the simulation cases from analytical results. The fluctuations observed are likely due to constraints of sample space for the random sampling.

Figure 3 Impact of number of walkers on node coverage: (a) absolute values and (b) relative differences with respect to models



(a)



(b)

5.2 Evaluation on object recall

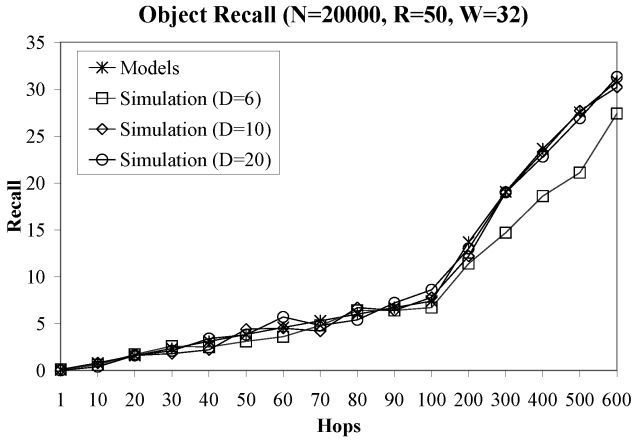
A query process can be considered as more efficient if a certain number of query messages yields higher object recall. In the following simulations, we inspect the influence of average node degree and replication ratio, respectively.

Effect of node degree. We simulated 32 random walkers searching for an object that has 50 randomly distributed replicas. The average node degree was varied as 6, 10, and 20. In each case, we used the average of the object recalls for ten searches. The simulation results are compared with the analytical models in Figure 4.

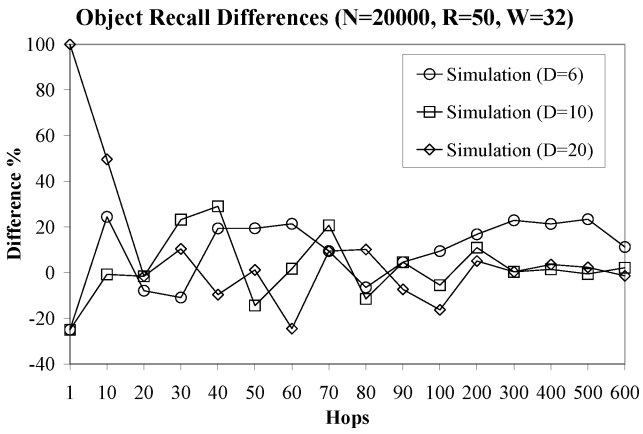
Note that the object recall for the analytical models are derived from node coverage. Since the actual recall value for a single search (run) heavily depends on the random choices made by each walker at each step, we expect certain fluctuation on the curves for the simulation results. (Ideally, the number of samples should be large enough to ensure stable results). Figure 4(a) indicates that the analytical models generate higher recall values than the simulations in general. The deviations of the simulation results from analytical models diminish as node degree increases. This is reasonable because the object recall is expected to increase as node coverage increases. Figure 4(b) shows the relative deviation from the analytical results. For a smaller number of hops, the object recall value obtained from simulations is too small to perform meaningful comparisons. As the recall increases with hop number, the effect of node degree becomes apparent – the higher the degree, the less the deviation from analytical models.

Effect of replication. We simulated a search in a graph with $d = 10$ and $w = 32$, while setting the replication ratio of the queried objects to 50 and 200, respectively. The results are plotted in Figure 5(a). The analytical values are somewhat similar to (but a little greater than) the simulation results (Figure 5(b)), barring some exceptions that are likely due to the limitation of sampling spaces. With both replication values, the relative difference tends to diminish as the hop number increases. This is because with more hops, the larger recall values recorded from each run produce more stable output than that for the case of fewer hops.

Figure 4 Impact of average node degree on object recall: (a) absolute values and (b) relative differences with respect to models

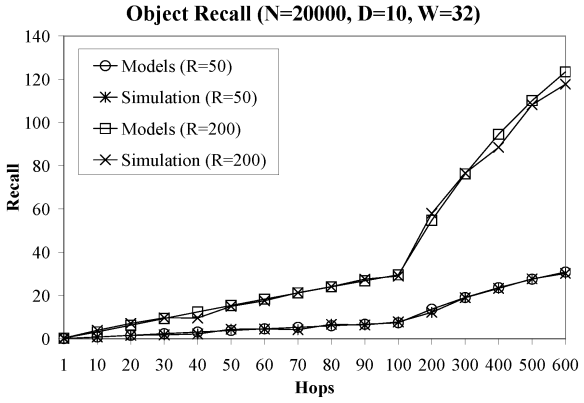


(a)

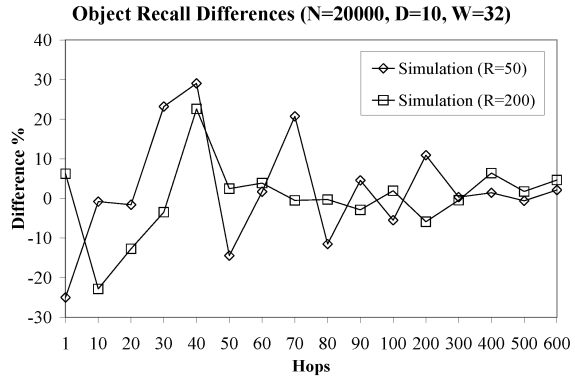


(b)

Figure 5 Impact of degree of replication on object recall: (a) absolute values and (b) relative differences with respect to models



(a)



(b)

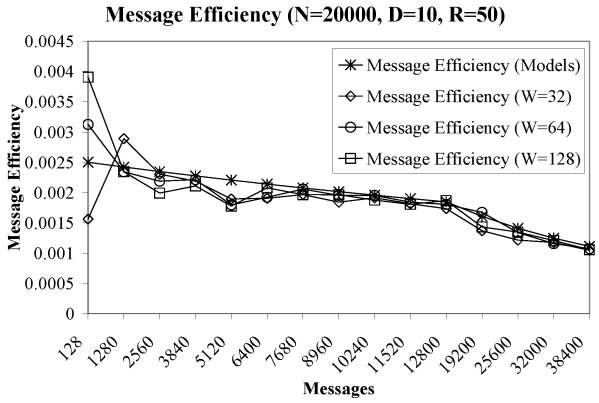
5.3 *Evaluation on message efficiency*

Message efficiency is a derived quantity: (object recall)/(message overhead). The expected object recall is in proportion to node coverage. According to equation (6), the efficiency should decrease as the message overhead increases. Since our analytic models provide an upper bound for expected node coverage, we also expect that the models give an upper bound for expected message efficiency. What interests us is to investigate how close the simulation results would approach the 'expected' upper bound, and what are the effects of search and topology parameters upon this approximation. The results indicate that it is reasonable to use the analytical results from our models as upper bounds of the expected message efficiency.

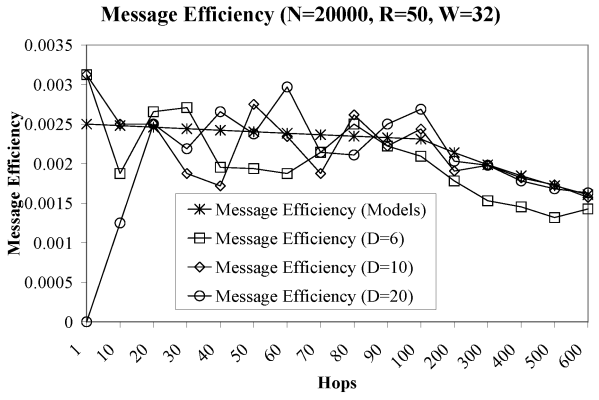
Effect of number of walkers. As the analytical models indicate, the number of walkers is seen to have no impact on the message efficiency. Figure 6(a) compares the message efficiency obtained from our models and also from simulation results for $w = 32$, $w = 64$, and $w = 128$, with 50 replicas per object. In our test cases, when the message overhead is low, the recall values obtained from the simulations are not stable enough for comparison. As the results smooth out with increasing messages, the analytical results tend to have better message efficiency than the simulation results. Also, the simulations suggest that the number of walkers has marginal impact on message efficiency. The relative differences between the analytical results and simulations are generally below 10% for all the three cases when message overhead >6400.

Effect of average node degree. The simulation results in Figure 6(b) show a similar trend as Figure 6(a) when we vary the average node degree while fixing the other parameters. As d has an impact on node coverage (see Section 5.1), the relative performance in terms of message efficiency with different d tends to be the same as the relative performance in terms of node coverage.

Figure 6 Impact on message efficiency: (a) impact of number of walkers and (b) impact of average node degree



(a)

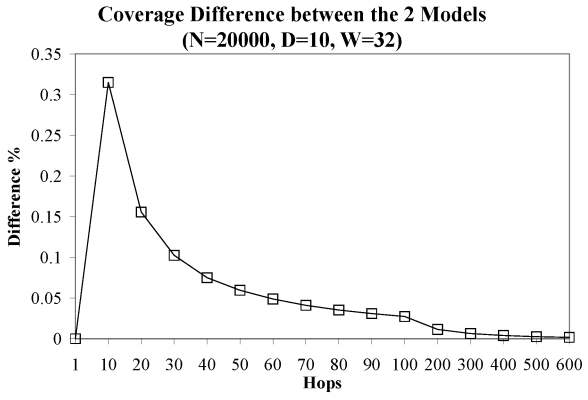


(b)

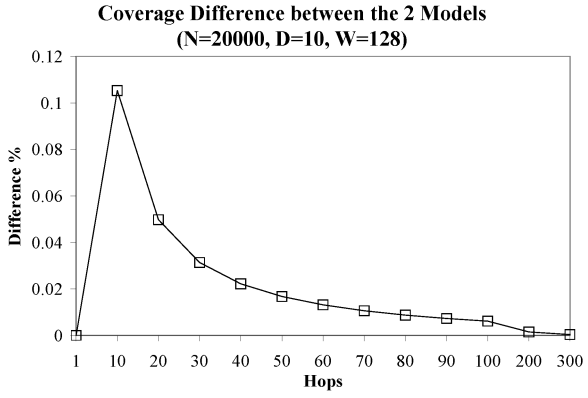
5.4 *Comparison of the two analytical models*

The two analytical models fit each other extremely well. The node coverage generated from either model are so close to each other that we represented the results from both as a single plot in the graphs so far. Figure 7 shows the relative difference in terms of node coverage, when $d = 10$ and w is 32 and 128, respectively. In both cases, the algebraic model generates slightly higher value of node coverage than the combinatorial model. Their differences are at most 0.3% for $w = 32$ and at most 0.1% for $w = 128$. Both models can serve as an upper bound for the estimate of node coverage and object recall.

Figure 7 Node coverage for the algebraic and combinatorial models: (a) 32 walkers and (b) 128 walkers



(a)



(b)

5.5 *Performance summary*

By comparing with simulation results, we evaluated the algebraic and the combinatorial models for computing Node Coverage (NC), Object Recall (OR), and Message Efficiency (ME) of unguided searches in random graphs. The impact of the average node degree, hop count, number of walkers, and replication ratios on the accuracy of the models, as studied via simulations, is summarised.

- The models give a little higher value of NC, OR, and ME than simulations.
- As d increases, the model's accuracy for NC and OR increases.
- As d increases (i.e., number of messages increases), accuracy of models for NC decreases.
- The NC is a function of the number of messages and appears independent of w.
- ME is seen to be almost independent of w.
- ME from the simulations becomes more stable as h increases.
- As d increases, the ME increases and accuracy of models also increases.

6 Random walk modelling and random sampling

The algebraic model simulates a process of random pick: there is a bag of N balls, at each step, we pick a ball and put it back. Then what is the expected number of distinct balls we will find after x attempts? Our algebraic model approximates this process with a derivative equation and the solution to this equation gives the formula for computing the node coverage.

The rationale of this formula can be validated by comparing the computation of success rate using the node coverage and that given by Bisnik and Abouzeid (2005).

Recall that the algebraic model gives the node coverage $u(x)$ in equation (3). The constant C is determined by the initial condition of a specific search process. If we take the following initial condition:

$$u(x) = 0 \mid x = 0 \tag{19}$$

and also use N to replace the term $N - 2$ in the exponent when N is large enough, we have the following:

$$u(x) = N(1 - e^{-x/N}). \tag{20}$$

By equation (18), the success rate of an object with r copies is then computed using node coverage:

$$p_{succ} = 1 - \left(1 - \frac{u(x)}{N}\right)^r = 1 - (e^{-x/N})^r. \tag{21}$$

On the other hand, (Bisnik and Abouzeid, 2005) suggests to compute the success rate using the following formula:

$$p_s = 1 - (1 - p)^{kT} \tag{22}$$

where $p = \frac{r}{N}$, $kT = x$, x is the message overhead, and k and T denote the number of walkers and the hop number, respectively, in Bisnik and Abouzeid (2005). To use the notations in our model, we rephrase this formula:

$$p_s = 1 - \left(1 - \frac{r}{N}\right)^x. \tag{23}$$

Also, equation (21) can be rephrased as:

$$p_{succ} = 1 - (e^{-r/N})^x. \tag{24}$$

Note that as N is large enough, we have:

$$\lim_{N \to +\infty} \left(1 - \frac{r}{N}\right)^N = e^{-r} \tag{25}$$

and thus,

$$\lim_{N \to +\infty} \left(1 - \frac{r}{N}\right) = \lim_{N \to +\infty} e^{-r/N} \tag{26}$$

and this equates the computation of p_{succ} in equation (21) and that of p_s in equation (22). These two formulas are equivalent as N goes to infinity.

Note that these two approaches come to different formulas with the same outcome, because they both simulate the same process of 'independent random picks' in which the proper computation for the success rate will always yield the same results.

However, the random walk in random graphs is not actually 'independent random picks', as a message forwarding is always associated with the status of the current node. The probability of finding a new node by the next message forwarding is decided not only by the current proportion of 'undetected' nodes in the network, but also affected by the current status of the sending node. This impact does not exist in the independent random pick process and it contributes to the differences between our algebraic model (combinatorial model also) and the simulation results.

7 Conclusions

In this paper, we provided two mathematical models to analyse the mechanism and performance of various search approaches in unstructured peer-to-peer networks. In our models, we use one type of random graph as the topology of the network and base our analysis on a probabilistic study of a set of concurrent processes, i.e., the forwarding of query messages. We assumed that when a query is in progress, different query messages are forwarded and processed independent of each other and our stochastic models grasp the key relation between the volume of query messages and the percentage of nodes visited by those messages, or node coverage. Based on this relation, we studied the performance of unguided search as a pseudo-random process in terms of various metrics. The results from our analysis models are compared with simulation results. We concluded that the analysis models are consistent and accurate. They approach the simulation results very closely in most of the overlay topology set ups.

Finally we compared the rationale of our models and the process of *random pick* and proved that these two processes are identical in terms of probabilities when the sampling space is large.

References

Bawa, M., Manku, G.S. and Raghavan, P. (2003) 'SETS: search enhanced by topic segmentation', *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.306–313.

Bisnik, N. and Abouzeid, A. (2005) 'Modeling and analysis of random walker search algorithm in P2P networks', *Second International Workshop on Hot Topics in Peer-to-Peer Systems*, pp.95–103.

Crespo, A. and Garcia-Molina, H. (2002) 'Routing indices for peer-to-peer systems', *Proceedings of the 22nd International Conference on Distributed Computing Systems(ICDCS'02)*, pp.23–34.

Eng, L.K., Crowcroft, J., Pias, M., Sharma, R. and Lim, S. (2005) 'A survey and comparison of peer-to-peer overlay network schemes', *IEEE Communications Survey and Tutorial*, Vol. 7, No. 2, pp.72–93.

Ge, Z., Figueiredo, D.R., Jaiswal, S., Kurose, J. and Towsley, D. (2003) 'Modeling peer-peer file sharing systems', *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies(INFOCOM 2003)*, Vol. 3, pp.2188–2198.

Gkantsidis, C., Mihail, M. and Saberi, A. (2006) 'Random walks in peer-to-peer networks: algorithms and evaluation', *Performance Evaluation*, Vol. 63, pp.241–263.

Liu, L., Ryu, K.D. and Lee, K. (2004) 'Supporting efficient keyword-based file search in peer-to-peer file share systems', *IEEE Global Telecommunications Conference (GLOBECOM '04)*, Vol. 2, pp.1259–1265.

Lv, Q., Cao, P., Cohen, E., Li, K. and Shenker, S. (2002) 'Search and replication in unstructured peer-to-peer networks', *Proceedings of the 16th International Conference on Supercomputing*, pp.84–95.

Reynolds, P. and Vahdat, A. (2003) 'Efficient peer-to-peer keyword searching', *Proceedings of the International Middleware Conference*, pp.21–40.

Risson, J. and Moors, T. (2006) 'Survey of research towards robust peer-to-peer networks: search methods', *Computer Networks*, Vol. 50, No. 17, December, pp.3485–3521.

Schmitz, C. (2004) 'Self-organizing a small world by topic', *Proceedings of the MobiQuitous'04 Workshop on Peer-to-Peer Knowledge Management*, 22 August, Boston, Massachusetts, USA.

Tang, C., Xu, Z. and Dwarkadas, S. (2003) 'Peer-to-peer information retrieval using self-organizing semantic overlay networks', *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp.175–186.

Tsoumakos, D. and Roussopoulos, N. (2003) 'Adaptive probabilistic search for peer-to-peer networks', *Proceedings of the 3rd International Conference on Peer-to-Peer Computing*, 1–3 September, Linkoping, Sweden, pp.102–109.

Wu, B. and Kshemkalyani, A.D. (2006a) 'Evaluation of analysis models for unguided search in unstructured P2P networks', *IFIP International Symposium on Network-Centric Ubiquitous Systems (NCUS 2006)*, LNCS 4097, Springer, 1–4 August, Seoul, Korea, pp.163–172.

Wu, B. and Kshemkalyani, A.D. (2006b) 'Analysis models for blind search in unstructured overlays', *5th IEEE Symposium on Network Computing and Applications (NCA)*, pp.223–226.

Zeinalipour-Yazti, D., Kalogeraki, V. and Gunopulos, D. (2004) 'On constructing internet-scale P2P information retrieval systems', *International Workshop on Databases, Information Systems and P2P Computing (DBISP2P 2004)*, LNCS 3367, pp.136–150.