Model-based Evaluation of Search Strategies in peer-to-peer Networks *

Rossano Gaeta and Matteo Sereno
Dipartimento di Informatica, Università di Torino,
Corso Svizzera, 185, 1049 Torino, Italia
email: {rossano, matteo}@di.unito.it

Abstract

This paper exploits a previously developed analytical modeling framework to compare several variations of the basic flooding search strategy in unstructured decentralized peer-to-peer (P2P) networks. The model predictions are used to compute system-oriented performance indexes (the average and the coefficient of variation of the number of query messages) as well as user-oriented measures (the probability of finding at least one replica of a resource, the average search time). The trade-off between the optimization of system-oriented measures and the improvement of user-oriented quality indexes is investigated for several variations of the basic flooding strategy suggesting that adding control parameters to the basic flooding mechanism might prove beneficial in this class of systems.

1. Introduction

The peer-to-peer (P2P) paradigm is becoming the basis for the development of several distributed networked services and applications. Besides the widely known file-sharing applications of (mostly) copyrighted audio and video files new P2P-based application have started to be adopted by Internet users, e.g., distributed grid computing [1], storage [7], web cache [13], Internet telephony [2], streaming [14, 25], conferencing [5], content distribution [17, 6]. Furthermore, file-sharing applications alone are responsible for close to 80% of the entire overall traffic in certain segments of the Internet [3].

The participants to the P2P-based application are termed as *peers* whose typical activity cycle consists of the location-retrieval-processing phases for a *resource*. Resources may have several *replicas* distributed among peers determining the resource *popularity*. In the so called *decentralized unstructured* P2P networks [15] each peer is responsible for maintaining an index of the resources that it is willing to share with the other peers in the network. The lack of global knowledge on which peers own a replica of a resource makes the process of locating a resource a rather complex task. We focus on the location phase of resources in decentralized unstructured P2P networks.

Locating a resource in this context can be done using blind or informed strategies. Roughly speaking, informed strategies keep a certain level of additional information of resource locations that speeds-up successive queries for similar objects. Blind methods are those where peers only have knowledge of other peers connected (at the application level) to them (*neighbors*) with whom they self-organize in an *overlay network* that forms the infrastructure on which search for resources takes place. Blind methods can be further divided in random walk based and *flooding* based strategies. In random walk based search strategies peers forward a query message to one randomly chosen neighbor at each step until a maximum threshold on the number of steps across the overlay network is reached. In the basic flooding search strategy a peer sends a resource-location request to all of its neighbors. This collection of neighbors may then forward the request to all of their neighbors (excluding, of course, the neighbor that sent the original request). These neighbors may then propagate the request to all their neighbors and so on up to a certain predefined maximum level. Hence, resource discovery is performed by flooding the network with resource-location request packets. We focus on blind, flooding based search strategies.

These search strategies must be carefully designed in order to achieve acceptable user satisfaction (a high *hit probability*) at a reasonable cost in terms of the traffic required to spread a query from an originator to other network participants that may own a replica of the requested resource. In this paper we use the analytical framework proposed in [10] to support the design of different search strategies that can be used in decentralized unstructured P2P networks.

The analytical framework that we use is based on the

* This work has been supported by the Italian Ministry for Education, University and Research (MIUR) within the framework of the Famous (PRIN) and Web-Minds (FIRB) projects.

1-4244-0054-6/06/$20.00

representation of the overlay network of a P2P-based application as a *generalized random graph* (GRG) [21]. A GRG models a snapshot of the topology of the P2P network where vertices represent peers and edges represent application-level connections between peers. GRGs are a generalization of the classical random graph model [9] that incorporates an arbitrary (but fixed) degree distribution; edges are selected independently and uniformly over the space of possible edges, constrained by the degree distribution. Although the connections among peers in the overlay network yield a constantly and randomly changing topology as the result of users joining and leaving the network, if we assume that the time scale of search operations (the objective of our study) is much shorter than the time scale of the P2P network topology evolution, we can reasonably assume that at any instant in time the snapshot of the P2P network topology can be viewed as an instance of a finite graph of size N.

The analysis is based on the derivation of the generating functions of the probability distribution of the number of query messages sent throughout the overlay network as well as the number of replica of a resource found during the search process. Well-known properties of generating functions are exploited to numerically compute parameters of the probability distribution that are used to evaluate the impact of the overlay network topology on the performance of flooding and to compare different variations of the basic flooding. The analysis of results suggests that improvement of the basic flooding mechanism can be obtained by introducing randomness in the forwarding of a query message to neighbors and to let this probability depend on the distance (expressed as the number of hops) from the query originator. The natural conclusion of this paper is that blind search methods that are based on flooding could be further improved by adding tunable parameters to more finely control the amount of overhead traffic required to locate resources.

The paper is organized as follows: Section 2 briefly sketches previous work on the analytical modeling of P2P networks that are related with our work. Section 3 summarizes the main results derived in [10] and defines the performance indexes we use to compare search strategies, Section 4 presents and analyzes results we obtained, and Section 5 draws conclusions and outlines future developments.

2. Related work

Besides [10] several papers have analyzed search strategies for unstructured decentralized P2P. The paper in [18] explores alternatives (expanding rings and random walks) to the classical flooding search strategies by means of simulation. The work in [12] exploits the theory of random graphs to prove properties of a generalization of the search that combines flooding and random walks. The work in [4] focuses on random walks and introduces a number of local search strategies that utilize high degree nodes in power-law graphs to reduce search time. The work in [11] quantifies the effectiveness of random walks for searching and construction of unstructured P2P networks. It also compares flooding and random walk by simulations on different network topologies. The authors of [23] introduce a scalable searching protocol for locating contents in random networks with heavy-tailed degree distributions. The analysis of the size of the giant connected component of a random graph with heavy tailed degree distributions under bond percolation is at the heart of their main results. The work in [16] proposes a strategy where peers build probabilistic routing tables that are used to forward search queries. Tables are constructed and maintained through the exchange of updated information among neighbors.

3. Random Graph Models of P2P Networks

The modeling approach proposed in [10] is based on the representation of the overlay networks of a P2P application as a generalized random graph (GRG). A GRG represents a family of graph instances with N vertices (nodes) where the degree of a randomly chosen node is specified by an arbitrary (but fixed) probability distribution. Results for GRGs are averages over the entire set of possible graph instances. In a collection of papers Newman developed a set of efficient tools for the analysis of GRGs (see for instance [20, 19, 21]). In particular, many interesting results are derived by using generating functions for probability distributions. In the following we briefly summarize the main results on the use of GRG to model the overlay networks of a P2P-based application. The interested reader may refer to [10] for full details.

Let $\{p_k\}$ be the probability distribution describing the number of neighbors of a randomly chosen peer. Its generating function is

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k.$$

We can also write the generating function of the degree distribution of the node reached by following one end of a randomly chosen edge (the starting edge is excluded) by

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)} = \frac{1}{z} G_0'(x),$$

where $G_0'(x)$ is the first derivative of $G_0(x)$ and $G_0'(1) = z$ is the average nodal degree of a randomly chosen node. It can be shown that the generating function of the probability distribution of the number of peers two hops away from a randomly chosen one is given by $G_0(G_1(x))$ where composition of generating functions can be further applied to obtain the generating function for the number of peers three hops away from a randomly chosen as $G_0(G_1(G_1(x)))$, and so on.

Assuming that number of nodes discovered at each step of the query spreading process are independent random variables[1], we can derive the generating function of the number of neighbors of a randomly chosen node up to a distance equal to TTL

$$\mathcal{G}(x,\mathrm{TTL}) = \left.\begin{array}{l} G_0(x)\times \\ \times G_0(G_1(x))\times \\ \times G_0(G_1(G_1(x)))\times \\ \dots \\ \times G_0(\underbrace{G_1(\dots G_1(x)\cdots)}_{\mathrm{TTL}-1}) \end{array}\right\} \text{TTL times}$$

3.1. Modeling the Search Strategies

In the flooding search mechanism a peer that is originating a search transmits the search message to each of its first neighbors and each peer receiving the search message forwards it to each of its neighbors. The propagation of the query continues until the *time to live* (TTL) of the query expires.

We consider variations of the basic mechanism where each peer (including the originator of the query) transmits the search message to one of its first neighbors with a given probability that can be a function of the distance (in terms of hops) between the peer and the query originator. A particular variation of the flooding search mechanism can be completely encoded in the function

$$g : I\!N \to [0,1].$$

This function represents the probability of forwarding a query as a function of the the distance between the peer and the query originator. This function allows to represent several variations of the flooding search strategy. In particular, the classical flooding can be represented by a constant function $g(d) = 1, \forall\, d \leq \mathrm{TTL}$ where TTL denotes the maximum time to live parameter.

If we denote as Q the random variable representing the number of query messages sent throughout the overlay network following the issuing of a query, it can be shown that the generating function for its probability distribution out to distance TTL is given by

$$\mathcal{Q}(x,g,\mathrm{TTL}) = \prod_{m=1}^{\mathrm{TTL}} Q_m(x,g), \qquad (1)$$

with $Q_1(x,g) = G_0(1+g(0)(x-1))$, and $Q_{h+1}(x,g) = Q_h(Q_1^{(h)}(x,g),g)$ for $h \geq 1$ where $Q_1^{(h)}(x,g) = G_1(1+g(h)(x-1))$.

1 In Section 3.5 we discuss about the impact of this assumption.

Starting from Equation (1) it is possible to define the average and the variance of the number of messages sent throughout the P2P network as

$$E[Q] = \mathcal{Q}'(1,g,\mathrm{TTL}) \quad \text{and}$$

$$Var[Q] = \mathcal{Q}''(1,g,\mathrm{TTL}) + \mathcal{Q}'(1,g,\mathrm{TTL})\left(1-\mathcal{Q}'(1,g,\mathrm{TTL})\right),$$

respectively. We also compute the coefficient of variation of the probability distribution as $C_x[Q] = \frac{\sqrt{Var[Q]}}{E[Q]}$.

3.2. Modeling the hit probability

We assume that the shared resources are uniformly distributed among the participant peers according to their popularity. In this manner we can derive an expression for the generating function for the number of neighbors of the peer that originates the query that have received a copy of the search message up to distance equal to TTL and have a replica of the resource. Let α be the probability that a randomly chosen peer holds a replica of the requested resource. It can be shown that the generating function for the total number of neighbors of a randomly chosen query originator peer that received a copy of the search message and holds a replica of the requested resource out to a distance TTL is

$$\mathcal{H}(x,g,\alpha,\mathrm{TTL}) = \prod_{m=1}^{\mathrm{TTL}} H_m(x,g,\alpha) \qquad (2)$$

where $H_h(x,g,\alpha) = Q_h(1+\alpha(x-1),g)$. From Equation (2) we can derive that the probability that a query is successful can be written as

$$p_{\text{hit}} = 1 - \mathcal{H}(0,g,\alpha,\mathrm{TTL}). \qquad (3)$$

3.3. Modeling the Search Time

We assume the knowledge[2] of the cumulative distribution functions (CDF) that a positive response from peers that are $1,2,3,\dots TTL$ hops away from the query originator returns to the query originator at a time less than or equal to t. We denote these distributions as $D_1(t)$, $D_2(t)$, $D_3(t)$, …, $D_{\mathrm{TTL}}(t)$.

It is possible to derive the (defective) CDF distribution of the time it takes to receive a least one positive response from any peer through level equal to TTL by time t as

$$D(t,g,\alpha,\mathrm{TTL}) = 1 - \mathcal{T}(0,g,\alpha,t,\mathrm{TTL}). \qquad (4)$$

with

$$\mathcal{T}(x,g,\alpha,t\mathrm{TTL}) = \prod_{h=1}^{\mathrm{TTL}} T_h(x,g,\alpha,t) \qquad (5)$$

2 The distribution $D_1(t)$, $D_2(t)$, $D_3(t)$, …, $D_{\mathrm{TTL}}(t)$ could be estimated by using measure based investigations or by using some auxiliary model.

and

$$T_h(x, g, \alpha, t) = H_h(1 + (x-1)D_h, g, \alpha).$$

As a particular case, search time can be expressed in terms of number of hops if we assume CDF $D_h(t)$ to be a unit step function that is time shifted by $h+1$ time units[3]. Starting from the CDF $D(t, g, \alpha, \text{TTL})$ we denote as $\overline{T}(g, \alpha, \text{TTL})$ the average time to obtain at least one positive reply in the case the resource is found, i.e., we excluded the time to obtain at least one reply when the resource is not found because in this case we should assume that the response time is either equal to some predefined timeout or equal to ∞.

3.4. Modeling the distance to the first hit

We can also define the distance (in number of hops) to the first peer that owns a replica of the resource. We have that $\mathcal{H}(0, g, \alpha, h)$ is the probability that the resource is not found in h hops and we define $\mathcal{F}(0, g, \alpha, h)$ as the probability that a search for the resource takes *exactly* h hops as

$$\mathcal{F}(0, g, \alpha, h) = \mathcal{H}(0, g, \alpha, h-1) - \mathcal{H}(0, g, \alpha, h).$$

The average hop distance from the first hit is thus

$$\overline{M}(g, \alpha, \text{TTL}) = \sum_{h=1}^{\text{TTL}} h \cdot \mathcal{F}(0, g, \alpha, h). \qquad (6)$$

3.5. Discussion on the Modeling Assumptions

Equations (1), (2), and (5) are developed with the simplifying assumption that the number of newly visited peers at each step of the flooding process are described by independent random variables. This allowed us to write the generating functions of the sum of random variables as a product of generating functions. But: how reasonable is this independence assumption? This assumption is equivalent to assume that the *clustering coefficient*[4] of the graph representing the overlay network approaches the value 0 as the size of the graph $N \to \infty$. A few measurement based studies found out that the measured value of the clustering coefficient of the overlay network of two popular P2P-based file-sharing applications, i.e., Gnutella [22] and the most recent Gnutella 2 [24], lies in the range $[0.018 - 0.020]$. This contradicts the assumption we did when deriving Equations (1), (2), and (5). Nevertheless, we compared the outcome of our model predictions with simulations on random graphs with tunable clustering coefficient: the main observation is that discrepancies between model and simulation are very small and become significative (about 15% error) only for $TTL > 4$. Since in real overlay networks (see again [22] and [24]) it has been found out that the vast majority of paths is of length less than or equal to 5 hops, we can conclude that the model we developed faithfully represents the behavior of search strategies even when overlay networks exhibit a (not very high) degree of clustering.

3 Here we assume that it takes h hops for the query to reach a peer and one hop for the positive reply to the query originator

4 It expresses the average probability that two neighbors of a node are neighbors themselves.

4. Experimental Results

The validation of the analytical framework summarized in Section 3 has been carried out in [10]; the model predictions were found to closely match simulation results obtained on instances of GRGs for several combinations of system parameters.

An efficient search strategy should keep the traffic overhead as low as possible while trying to achieve high hit probabilities and low search times. In this section we use the GRG framework to investigate the following issues:

- effects of the overlay topology characteristics on the performance of the basic flooding mechanism;
- evaluation of different flooding-based search strategies on power-law overlay networks.

4.1. Overlay network topology and performance of flooding

In the first set of experiments we evaluate the characteristics of the basic flooding strategy for three topologically different P2P networks. In particular, we consider the degree distribution of the GRG representing the overlay network to be a power-law with exponential cutoff, Poisson, or uniform. Formally, we consider the following expressions for p_k whose parameter we fit to obtain an average degree equal to 3.5 in all the three cases:

- $p_k = \frac{k^{-\tau} e^{-k/\kappa}}{Li_\tau(e^{-1/\kappa})}$ where τ, κ are the power-law exponent and cutoff, respectively. The notation $Li_n(x)$ is used to express the n-th poly-logarithm function, i.e., $Li_n(x) = \sum_{k=1}^{\infty} \frac{x^k}{k^n}$;
- $p_k = \frac{e^{-z} z^k}{k!}$ for the Poisson distribution (z is the parameter of the Poisson distribution and is equal to average degree);
- $p_k = \frac{1}{b-a}$ where a and b are the lower and the upper bound of the uniform distribution, respectively

The basic flooding strategy is modeled by considering the constant function $g(d) = 1$, for $d = 1, 2, \ldots, \text{TTL}$.

Table 1 reports the values of the average number of query messages as well as the coefficient of variation of the probability distribution of random variable Q for $\text{TTL} = 4$. Furthermore, it reports the values of p_{hit} for increasing values of the resource popularity α. It can be noted that despite the fact that all three degree distributions share the same average value, the overall average number of messages ($E[Q]$) differs orders of magnitude. In particular, for the power-law overlay topology we observe a huge average amount of overhead traffic with respect to Poisson and uniform degree distributions. This is a well known phenomenon that can be intuitively explained by the observation that starting from a randomly chosen node the probability of encountering a high degree node (a *hub* in the common terminology) as the query travels from the originator becomes higher. Once a very high degree node receives a query message soon the query propagates to a very large number of intermediate forwarders that will spread the query to a large fraction of the participants. This observation also explains the observed values of the coefficient of variation $C_x[Q]$. Both observation are a direct consequence of the extremely high variance of a probability distribution defined as a power-law with exponential cutoff. Also, we observe that the hit probabilities increases as the resource popularity increases but for a fixed value of α we observe remarkable differences for the different degree distributions. Again, the very high value of the hit probability in the case of power-law overlay topologies is explained by the fact that a huge amount peers is contacted to check for the resource and hence a high probability of finding at least one replica is obtained.

But why are power-law topologies so important? It has been shown by several authors that the measured degree distributions of two highly popular P2P-based file-sharing applications, i.e., Gnutella [22] and the most recent Gnutella 2 [24], exhibit high variance. Power-law probability distributions are one possible mathematical description of such heavy-tailed distributions and they deserve a closer examinations.[5]

4.2. Comparison of variations of the basic flooding strategy

The basic flooding strategy suffers from poor scalability and granularity. By granularity we mean that the performance of the mechanism is dictated by only one parameter, i.e., the time-to-live. A minimum increase of the value of the TTL yields an exponential increase in the number of messages sent throughout the overlay network thus leading to poor scalability. The basic mechanism could be improved by adding control parameters whose values could be tuned to meet pre-specified performance requirements on both the average number of messages originating from a request and the hit probability. We make a first attempt in this section to devise search strategies where two additional control parameter are available for tuning their performance. To this end, we consider two families of search strategies defined as:

$$g_1(d, p_{fwd}, \beta) = p_{fwd}(d+1)^{-\beta}$$

and

$$g_2(d, p_{fwd}, \beta) = \begin{cases} 1 & \text{if } d = 0 \text{ or } d = 1 \\ p_{fwd}(d+1)^{-\beta} & \text{otherwise.} \end{cases}$$

Both families are defined by parameters p_{fwd} and β whose values control the decaying structure of the forwarding probability that peers use to cooperate in the query spreading process. The strategy defined by $g1$ is a distance-dependent probabilistic flooding while strategy defined by $g2$ represents a hybrid scheme where an initial shallow flooding with depth equal to two hops is followed by a distance-dependent probabilistic flooding for larger distances from the query originator. Here the basic flooding strategy is modeled by $g1(d, 1.0, 0.0) = 1.0,\ \forall d$.

To prove that the addition of control parameters is crucial in obtaining better performance we consider a rather uncommon resource whose popularity is $\alpha = 0.0001$. Comparison must be made in these rather tough cases since searching for very common resources requires no efforts to define clever search strategies: any blind strategy, e.g., basic flooding with low TTL, is able to provide high hit probabilities at a low cost in terms of average number of messages. We consider a power-law with exponential cutoff degree distribution for the overlay network whose parameters are those used in Section 4.1 and we limit the scope of the search strategies to four hops, i.e., TTL=4.

The setting we consider is the following: we consider the problem of locating the resource and want to design our search strategy in such a way that the resulting p_{hit} value is equal to 0.5. The model solution for the basic flooding mechanism yields $p_{hit} = 0.36$ for TTL=3 and

	Power-law $\tau = 2.041289, \kappa = 500$ $E[Q] = 1,216,464.12$ $C_x[Q] = 4.52$	Poisson $z = 3.5$ $E[Q] = 190.35$ $C_x[Q] = 0.53$	Uniform Unif$[2,5]$ $E[Q] = 123.71$ $C_x[Q] = 0.28$
α	p_{hit}	p_{hit}	p_{hit}
0.0001	0.8500241	0.1342200	0.0122889
0.001	0.9460802	0.2835374	0.1158808
0.01	0.9820292	0.8521187	0.6940224
0.1	0.9948324	0.9997804	0.9998434

Table 1. Comparisons among three different network topologies with power-law (with $\tau = 2.04128906$, $\kappa = 500$), Poisson (with $z = 3.5$), and uniform (with $a = 2$ and $b = 5$) distributions for TTL=4

5 The importance of power-law degree distributions has already been proved in several different scientific fields ranging from biology to social science, from economy to epidemiology. One source of details for this topic is [8].

	$g1(d, 1.0, 0.0)$	$g1(d, 0.562, 0.0)$	$g1(d, 0.925, 1.0)$	$g2(d, 0.152, 0.0)$	g2(d,0.510,1.0)
$E[Q]$	1,216,464	122,655	38,918	30,536	29,455
$C_x[Q]$	4.52	4.63	4.33	4.17	4.11
$\overline{T}(g, \alpha, \text{TTL})$	0.96	0.99	0.99	0.98	0.97

	$g1(d, 1.0, 0.0)$	$g1(d, 0.848, 0.0)$	$g2(d, 0.462, 0.0)$
$E[Q]$	1,216,464	630,368	263,929
$C_x[Q]$	4.52	4.54	4.44
$\overline{T}(g, \alpha, \text{TTL})$	0.96	0.97	0.97

Table 2. Comparison among different flooding based search strategies on network topology with a power law with $\tau = 2.041289$ and $\kappa = 500$. Resource popularity is $\alpha = 0.0001$, TTL=4, and the design requirement on the p_{hit} value is equal to 0.5 (upper table) and $p_{hit} = 0.75$ (lower table). The tables report the values for $E[Q]$, and $C_x[Q]$

$p_{hit} = 0.85$ for TTL=4. The average number of search messages is $17,337$ and $1,216,464$, respectively. Here we face the poor granularity problem of the simple flooding. We simply are not able to design a search strategy to meet the pre-defined requirement. Table 2 (upper table) reports the values of p_{fwd} and β for both strategy $g1$ and $g2$ that yield $p_{hit} = 0.5$: the table shows the corresponding values of $E[Q]$ and $C_x[Q]$.

It can be noted that by tuning p_{fwd} and β it is possible to meet the design requirement and to reduce the average number of messages sent throughout the network. Both $g1$ and $g2$ search strategies are effective but $g2$ yields the best performance. It should be pointed out that neither strategies are able to provide $p_{hit} = 0.5$ for $\beta = 2$. Slight improvements can still be obtained by letting β assume non integer values; in this case, the maximum value for $g1$ is $\beta = 1.1$ while for $g2$ we obtained $\beta = 1.5$.

Is there any hidden cost behind this sharp reduction of the average number of messages? Table 2 (upper table) also presents the conditional average time to obtain at least one response. It can be noted that this performance index is almost insensitive to the search strategy therefore we can conclude that there is no penalty for the application users in adopting the proposed alternative strategies in terms of increased response time.

As a further check, we considered the behavior of the search strategies $g1$ and $g2$ on a more stringent design goal. In particular, we considered a target $p_{hit} = 0.75$ and repeated a similar analysis. The results are presented in Table 2 (lower table). It can be noted that both $g1$ and $g2$ achieve the desired value for the hit probability only for $\beta = 0$. Also in this case the $g2$ policy performs better with a reduction of the average number of query messages that is almost one order of magnitude with respect to the basic flooding. Again, no extra costs in terms of response time are paid by applications users.

One may hypothesize that the reduction of the average number of messages is due to the particular choice of the parameters of the power-law degree distribution and of the TTL. We conducted the same experiments on a different power-law topology whose parameters are $\tau = 2.1$ and $\kappa = 125$ and considered TTL=5. Tables 3 (lower and upper tables) summarize the results we obtained. The results confirm that for the target value $p_{hit} = 0.5$ one order of magnitude reduction of the average number of query messages can be achieved by both $g1$ and $g2$ strategies. For $p_{hit} = 0.75$ the average number of messages is more than halved by using strategy $g2$ that performs better than $g1$ in both cases for the target p_{hit} values.

To verify that the gain that is obtained by adding control parameters to the basic flooding is not dependent on the functional form of the degree distribution we considered both Poisson and uniform degree distributions for representing the overlay network topology. We set the average degree to 15 and repeated the experiments whose results are presented in Tables 4 (upper and lower tables for Poisson nodal degree distribution) and Tables 5 (upper and lower tables for uniform nodal degree distribution). In both cases the conclusions we drawn for the power-law topology still hold. Strategies $g1$ and $g2$ effectively decrease $E[Q]$ at no additional cost in terms of response time.

5. Conclusions and future work

In this paper we exploited an analytical framework we developed to analyze the performance of flooding-based search strategies in unstructured decentralized P2P networks. The model is based on the use of generalized random graphs to represent the overlay network of the P2P-based application. The generating function of several probability distributions are derived under the hypothesis that the number of newly visited peers at each step of the flooding process are described by independent random variables. The validity of this assumption has been discussed and the model has been used to evaluate the impact of introducing additional control parameters on the performance of search strategies both from a system-oriented viewpoint and from a user-oriented viewpoint. In particular, we proved that the

	$g1(d, 1.0, 0.0)$	$g1(d, 0.65, 0.0)$	$g2(d, 0.37, 0.0)$
$E[Q]$	397,892	47,552	22,194
$C_x[Q]$	2.82	2.95	2.58
$\overline{T}(g, \alpha, \text{TTL})$	1.12	1.16	1.16

	$g1(d, 1.0, 0.0)$	$g1(d, 0.895, 0.0)$	$g2(d, 0.71, 0.0)$
$E[Q]$	397,892	229,955	145,626
$C_x[Q]$	2.82	2.85	2.76
$\overline{T}(g, \alpha, \text{TTL})$	1.12	1.13	1.14

Table 3. Comparison among different flooding based search strategies on network topology with a power law with $\tau = 2.1$ and $\kappa = 125$. Resource popularity is $\alpha = 0.0001$, TTL=5, and the design requirement on the p_{hit} value is equal to 0.5 (upper table) and $p_{hit} = 0.75$ (lower table). The tables report the values for $E[Q]$, and $C_x[Q]$

	$g1(d, 1.0, 0.0)$	$g1(d, 0.596, 0.0)$	$g2(d, 0.335, 0.0)$
$E[Q]$	54,239	7,191	7,052
$C_x[Q]$	0.25	0.32	0.22
$\overline{T}(g, \alpha, \text{TTL})$	0.98	1.01	0.99

	$g1(d, 1.0, 0.0)$	$g1(d, 0.717, 0.0)$	$g2(d, 0.497, 0.0)$
$E[Q]$	54,239	14,749	14,422
$C_x[Q]$	0.25	0.29	0.23
$\overline{T}(g, \alpha, \text{TTL})$	0.98	1.01	0.99

Table 4. Comparison among different flooding based search strategies on network topology with a Poisson degree distribution with $z = 15$. Resource popularity is $\alpha = 0.0001$, TTL=4, and the design requirement on the p_{hit} value is equal to 0.5 (upper table) and $p_{hit} = 0.75$ (lower table). The tables report the values for $E[Q]$, and $C_x[Q]$

average number of messages sent throughout the overlay network can be reduced by two orders of magnitude for two families of distance-dependent probabilistic flooding. We conclude that granularity of the basic flooding strategy can be increased and effective strategies can be devised and analyzed by means of the framework we defined. A natural consequence of this work is to ask: can we define additional control parameters to improve the performance of flooding-based strategies? If yes, which ones? We are investigating the answer to the above questions. It is likely that exploiting the heterogeneity of P2P-based applications, as already suggested by several authors, can lead to improvement of blind search strategies. As for the case of random walk based search strategies, additional control parameters could be the degree of nodes that cooperate in the spreading process of the query throughout the overlay network.

References

[1] Entropia. Technical report. http://www.entropia.com.

[2] Global Index (GI). Technical report. URL http://www.skype.com/skype_p2pexplained.htm.

[3] The true picture of peer-to-peer file sharing. Technical report, Cachelogic Research, July 2004. URL http://www.cachelogic.com/research/.

[4] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in Power-Law Networks. *Physical Review E*, 64, 2001.

[5] Y.-H. Chu, S. Rao, , and H. Zhang. A case for end system multicast. *IEEE Journal on Selected Areas in Communications*, 20(8), 2002. Special Issue on Networking Support for Multicast.

[6] B. Cohen. BitTorrent protocol specification. In *First Workshop on Economics of Peer-to-Peer Systems (P2P '03)*, Berkeley, CA, USA, June 2003.

[7] F. Dabek, M. F. Kaashoek, D. Karger, R. Morris, and I. Stoica. Wide-Area Cooperative Storage with CFS. In *Proc. of the 18-th ACM Symposium on Operating Systems Principles (SOSP '01)*, Banff, Canada, 2001.

[8] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks (From Biological Nets to the Internet and WWW)*. Oxford University Press, 2004.

[9] P. Erdõs and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6, 1959.

[10] R. Gaeta, G. Balbo, S. Bruell, M. Gribaudo, and M. Sereno. A simple analytical framework to analyze search strategies in large-scale peer-to-peer networks. *Performance Evaluation*,

	$g1(d, 1.0, 0.0)$	$g1(d, 0.522, 0.0)$	$g2(d, 0.249, 0.0)$	$g2(d, 0.845, 1.0)$
$E[Q]$	103,086	8,071	7,645	7,572
$C_x[Q]$	0.54	0.57	0.46	0.45
$\overline{T}(g, \alpha, \mathrm{TTL})$	0.96	1.01	0.98	0.97

	$g1(d, 1.0, 0.0)$	$g1(d, 0.646, 0.0)$	$g2(d, 0.392, 0.0)$
$E[Q]$	103,086	18,526	17,335
$C_x[Q]$	0.54	0.56	0.49
$\overline{T}(g, \alpha, \mathrm{TTL})$	0.96	1.01	0.99

Table 5. Comparison among different flooding based search strategies on network topology with a unifor degree distribution between $[1, 29]$ (average equal to 15). Resource popularity is $\alpha = 0.0001$, TTL=4, and the design requirement on the p_{hit} value is equal to 0.5 (upper table) and $p_{hit} = 0.75$ (lower table). The tables report the values for $E[Q]$, and $C_x[Q]$

62(1–4):1–16, 2005. Proceedings of the conference *Performance 2005*.

[11] C. Gkantsidis, M. Mihail, and A. Saberi. Random Walks in Peer-to-Peer Networks. In *IEEE INFOCOM 2004*, Hong Kong, China, Mar 2004.

[12] C. Gkantsidis, M. Mihail, and A. Saberi. Hybrid Seach Schemes for Unstructuired Peer-to-Peer Networks. In *Proc. IEEE Infocom 2005*, Miami, FL, USA, 2005. IEEE Comp. Soc. Press.

[13] S. Iyer, P. Rowstron, and P. Druschel. Squirrel: a Decentralized Peer-to-Peer Web Cache. In *Proc. of ACM Symposium on Principles of Distributed Computing (PODC '02)*, Monterey, CA, USA, 2002.

[14] S. Iyer, P. Rowstron, and P. Druschel. SplitStream: High-Bandwidth Multicast in Cooperative Environments. In *Proc. of the 19-th ACM Symposium on Operating Systems Principles*, Bolton Landing, NY, USA, 2003.

[15] G. Kan. Chapter 8: Gnutella. In A. Oram, editor, *Peer-to-Peer Harnessing the Power of Disruptive Technologies*. O'Reilly, 2001.

[16] A. Kumar, J. Xu, and E. W. Zegura. Efficient and Scalable Query Routing for Unstructured Peer-to-Peer Networks. In *Proc. IEEE Infocom 2005*, Miami, FL, USA, 2005. IEEE Comp. Soc. Press.

[17] J. Li, P. A. Chou, and C. Zhang. Mutualcast: An Efficient Mechanism for Content Distribution in a Peer-to-Peer (P2P) Network. Technical report, Microsoft Research, MSR-TR-2004-100, 2004.

[18] Q. Lv, E. Cohen, K. Li, and S. Shenker. Search and Replication in unstructured peer-to-peer network. In *International Conference on Supercomputing (ICS02)*, 2002.

[19] M. E. J. Newman. The spread of epidemic disease on networks. *Physical Review E*, 66, 2002.

[20] M. E. J. Newman. The structure and function of networks. *Computer Physics Communications*, 147:40–45, 2002.

[21] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64, 2001.

[22] M. Ripeanu. Peer-to-Peer Architecture Case Study: Gnutella Network. In *In proceedings of IEEE 1st International Conference on Peer-to-peer Computing (P2P2001)*, Linkoping Sweden, Aug 2001.

[23] N. Sarshar, V. Roychowdury, and P.O. Boykin. Percolation search algorithm, making unstructured p2p networks scalable. In *Proc. of the Fourth IEEE P2P 2004*, Zurich, Switzerland, 2004. IEEE Computer Society Press.

[24] D. Stutzbach, R. Rejaie, and S. Sen. Characterizing Unstructured Overlay Topologies in Modern P2P File-Sharing Systems. In *Proc. of the ACM SIGCOMM Internet Measurement Conference*, Berkeley, CA, USA, 2005.

[25] X. Zhang, J. Liu, B. Li, and T. P. Yum. DONet/CoolStreaming: A Data-driven Overlay Network for Live Media Streaming. In *Proc. IEEE Infocom 2005*, Miami, FL, USA, 2005. IEEE Comp. Soc. Press.