Link Lifetimes and Randomized Neighbor Selection in DHTs

Zhongmei Yao and Dmitri Loguinov*
Department of Computer Science, Texas A&M University
College Station, TX 77843 USA
{mayyao, dmitri}@cs.tamu.edu

***Abstract*—Several models of user churn, resilience, and link lifetime have recently appeared in the literature [12], [13], [34], [35]; however, these results do not directly apply to classical Distributed Hash Tables (DHTs) in which neighbor replacement occurs not only when current users die, but also when new user arrive into the system, and where replacement choices are often restricted to the successor of the failed zone in the DHT space. To understand neighbor churn in such networks, this paper proposes a simple, yet accurate, model for capturing link dynamics in structured P2P systems and obtains the distribution of link lifetimes for fairly generic DHTs. Similar to [8], our results show that deterministic networks (e.g., Chord [28], CAN [24]) unfortunately do not extract much benefit from heavy-tailed user lifetimes since link durations are dominated by small remaining lifetimes of newly arriving users that replace the more reliable existing neighbors. We also examine link lifetimes in randomized DHTs equipped with multiple choices for each link and show that users in such systems should prefer neighbors with *smaller zones* rather than larger age as suggested in prior work [13], [30]. We finish the paper by demonstrating the effectiveness of the proposed min-zone neighbor selection for heavy-tailed user lifetime distributions with the shape parameter α obtained from recent measurements [4], [31].**

I. Introduction

Resilience of distributed peer-to-peer (P2P) networks under user churn has recently attracted significant attention and has become an important research area [11], [12], [17], [25], [34]. Traditional metrics of performance in this analysis have been the ability of the graph to stay connected during user departure [13], [17], [23], behavior of immediate neighbors during churn [11], data delivery ratio [30], evolution of out-degree [12] and in-degree [34], and churn rate in the set of participating nodes [8]. All metrics above depend on one fundamental parameter of churn – *link lifetime*, which is defined as the delay between formation of a link and its disconnection due to a sudden departure of the adjacent neighbor.

In many P2P networks, each user v creates k links to other peers when joining the system, where k may be a constant or a function of system size [18], and detects/repairs failed links in order to remain connected and perform P2P tasks (e.g., routing and key lookups) [24], [25], [26], [28]. Under fairly general conditions on user lifetimes [12], [34], link behavior is often modeled as an ON/OFF process in which each link is either ON at time t, which means that the corresponding user is currently alive, or OFF, which means that the user adjacent to the link has departed from the system and its failure is in the process of being detected and repaired. ON durations of links are commonly called *link lifetimes* and their OFF durations are called *repair delays*. With this setup, it is not hard to see that link lifetimes play a key role in the study of resilience, performance, and reliability of P2P networks. For instance, longer average link lifetime means that users must repair failed links less frequently, which leads to smaller churn rates in the terminology of [8], and queries are less likely to encounter dead neighbors during routing [11], which yields larger data delivery ratios [30] and higher lookup success rates.

If links do not switch to other users during each ON duration (i.e., keep connecting to the same neighbors until they fail), then link durations are simply *residual lifetimes* of original neighbors. We call this model *non-switching* and note that it applies to certain unstructured P2P networks [7] and some DHTs [21]. Link lifetimes for non-switching systems have been studied in fair detail under both age-independent [12], [34] and age-biased [30], [35] selection. However, many DHTs actively switch links to new neighbors before the current neighbor dies in order to balance the load and ensure DHT consistency. We call such systems *switching* and note that their link lifetimes require entirely different modeling techniques, which we present below.[1]

A. Analysis of Existing DHTs

We start by introducing a stochastic process that keeps track of the changes in the identity of neighbors adjacent to the i-th link of a given user v as the system experiences churn. We show that this process is a regular semi-Markov chain whose first hitting time to the absorbing state corresponding to the failure of the last user holding the link is link lifetime R. Using this model, we find that the distribution of R is determined not only by lifetimes of attached users, but also by the zone size of the original neighbor holding the link.

We next obtain the Laplace transform of the distribution of R and derive its expected value $E[R]$ for general user lifetimes L, including heavy-tailed cases. We then use this result to show that in systems with exponential peer lifetimes, link

*Supported by NSF grants CCR-0306246, ANI-0312461, CNS-0434940, CNS-0519442, and CNS-0720571.

[1] In the notation of [8], switching/non-switching are agnostic neighbor replacement strategies, where the former is called Active Preference List (APL) and the latter encompasses both Passive Preference List (PPL) and Random Replacement (RR).

lifetime R follows the same exponential distribution, which indicates that for such cases link lifetimes are very similar to those in networks without switching [12]. However, for heavy-tailed peer lifetimes (e.g., Pareto) observed in many real P2P networks [4], [27], [31], our model of link lifetime R shows that R is stochastically *smaller* than the residual lifetime Z of the initial neighbor holding the link and, as first observed in [9], the mean link lifetime $E[R]$ is very close to $E[L]$. This is in stark contrast to the results of [12] where $E[R]$ is several times larger than $E[L]$ depending on Pareto shape α of the lifetime distribution (e.g., $E[R] \approx 11.1E[L]$ for $\alpha = 1.09$ observed in [31] and $E[R] \approx 16.6E[L]$ for $\alpha = 1.06$ observed in [4]). This phenomenon occurs because older (i.e., more reliable) neighbors in DHTs are replaced with new arrivals that exhibit much shorter remaining lifetimes. As a result, classical DHTs unfortunately do not extract any benefits from heavy-tailed user lifetimes and suffer much higher link churn rates than the corresponding unstructured systems [12]. A similar conclusion was obtained in [8] for query failure rates in Chord.

B. Improvements

One method of overcoming the problem identified above is to utilize randomized DHTs (e.g., randomized Chord [10], randomized hypercube [20], and Symphony [19]) in which the i-th finger pointer of a given user v is randomly selected from some set S_i of possible locations in the DHT space. By trying multiple options in S_i and linking to the user with the best characteristics (which we determine below), the hope is to improve link lifetime and reduce the impact of churn on system performance.[2] The first randomized technique, which we call *max-age*, selects m points in S_i uniformly randomly and connects v to the user with the largest age (this method was suggested in [30] for DHTs and [35] for unstructured P2P systems). While quite effective in non-switching scenarios, this strategy has minimal impact in DHTs since link lifetime is determined by the remaining session length of not the *first*, but the *last* neighbor holding the link.

To overcome this limitation, we propose a novel randomized strategy that stems from our model of link lifetime R. Our theoretical results show that neighbors with larger zones (e.g., in Chord [28], this means larger distance to the predecessor) are less reliable as they are more likely to be hit by a new arrival whose remaining lifetime will be small. To extract benefits from randomized selection, we show that users must prefer neighbors in S_i with the *smallest zone size* rather than maximum age or any other characteristic. We call this strategy *min-zone* and show that it is vastly more effective than max-age selection given lifetime distributions observed in real systems [4], [31]. In addition to reduced link churn, min-zone selection benefits DHTs by balancing the load such that users with smaller zone sizes are responsible for fewer keys while forwarding more queries.

Note that min-zone selection allows one to achieve a spectrum of neighbor-selection strategies, where $m = 1$ corresponds to regular switching behavior of DHTs and $m \to \infty$ (assuming $|S_i| \to \infty$) corresponds to a non-switching system (in fact, different links of the same peer may use different m depending on the size of each S_i). However, unlike purely non-switching networks that create inconsistences in finger tables and sometimes require routing along successor/predecessor links, min-zone selection always keeps the network consistent.

We finish the paper by showing that under min-zone selection and shape parameter $1 < \alpha \leq 2$, the mean link lifetime $E[R]$ tends to infinity as the number of samples m becomes large. We also suggest simple formulas for $E[R]$ using examples of Pareto shape α obtained from recent measurements [4], [31] and show simple results demonstrating the growth rate of $E[R]$ as a function of m.

The rest of the paper is organized as follows. Section II overviews related work. Section III introduces our model of user churn, DHT space, and zone splitting. In Section IV, we propose a general model of analyzing link lifetimes based on the semi-Markov chain associated with neighbor zone occupancy. We then apply this model to examine link lifetimes in deterministic DHTs in Section V and randomized DHTs in Section VI. Section VII concludes the paper.

II. Related Work

Among the recent studies of link lifetimes, one direction focuses on non-switching P2P systems. Leonard *et al.* [12] show that heavy-tailed lifetimes allow link lifetime $E[R]$ to be significantly larger than user lifetime $E[L]$. Additional results of this model and its application to unstructured networks are available in [13], [34], [35]. Another recent study [30] examines DHTs without switching with a focus on the *delivery ratio*, which is the fraction of time that all forwarding nodes between each source and destination are alive. Their results show that the delivery ratio is a function of link lifetime R for all examined neighbor-selection techniques.

The other direction covers switching networks exemplified by traditional DHTs. Godfrey *et al.* [8] study the impact of node-selection techniques on the churn rate and observe that switching DHTs exhibit dramatically smaller link lifetimes than non-switching networks. Krishnamurthy *et al.* [11] compute the probability that neighbors in Chord are in one of three states (alive, failed, or incorrect) and use this model to predict lookup consistency and latency.

Additional work [2], [5], [14], [15], [16], [25], [29] focuses on measurement and simulation of structured P2P systems under churn.

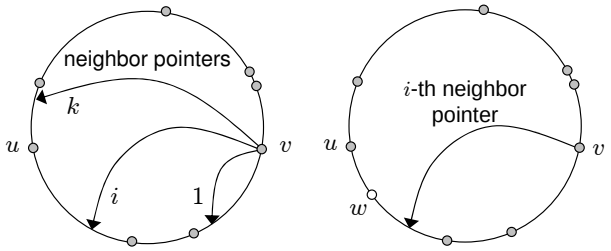
III. General DHT Model

We start by formulating assumptions on the DHT space, churn model, and link switching in DHTs.

A. Assumptions

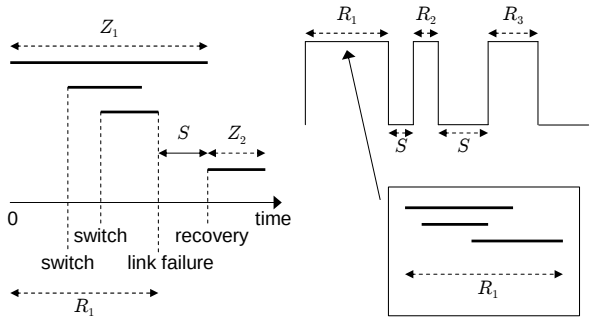
Without loss of generality, we assume that the network maps keys and users into the same identifier (ID) space, which is a continuous ring in the interval $[0, 1)$ [22]. Each user is responsible for a fraction of the DHT space from itself to its

[2]Note that this method only works when set S_i is sufficiently large. We assume that each node has at least one link that satisfies this condition.



(a) k links of user v (b) arriving node w

Fig. 1. User v's neighbors in the DHT.



(a) one recovery cycle (b) ON/OFF link behavior

Fig. 2. The i-th link failure and replacement of user v who joins at time 0 in a DHT, $1 \leq i \leq k$.

successor, which we call the user's zone. To facilitate routing, each joining peer v selects and then monitors using some stabilization technique k links in the DHT space as shown in Fig. 1(a).

For the churn model, we adopt the recently introduced [34] framework of n alternating renewal processes representing periodic online/offline behavior of users observed in real traces [8], [31]. In this model, each user i is viewed as alternating between online and offline states, where the duration of each state is random and has some user-specific distribution. While the total number of users n is fixed, the number of *currently alive* peers N_t at time t is a random process that fluctuates over time. Once stationarity is reached, we usually replace N_t with its limiting version $N = \lim_{t\to\infty} N_t$. We finally assume that when a particular user rejoins the system, it generates a new random ID (e.g., based on its IP-port pair) instead of using the same fixed hash. Note that the use of new IDs helps balance the load in the DHT [28], [32]. As a consequence of this churn model [34, Theorem 5], user arrivals into the system follow a Poisson process with a constant rate $\lambda = E[N]/E[L]$, where $E[N]$ is the average number of users in the steady state and $E[L]$ is the mean user lifetime.

B. Neighbor Dynamics

Note that the main focus of the paper is on the behavior of one particular link i in Fig. 1(a) and neighbors adjacent to it during v's online session. As user v continues to stay in the system, the identity of its neighbors (i.e., successors of its neighbor pointers) may change over time as users join and leave the system. There are two types of changes in neighbor tables – graceful handoffs of an existing zone to another user and abrupt departures without explicit notification of v. The former type, which we call a *switch*, occurs when a new arrival takes ownership of a link by becoming the new successor of the corresponding neighbor pointer. This is shown in Fig. 1(b) where a new arrival w splits the zone of an existing neighbor u and becomes the new neighbor of v. The latter type of neighbor change, which we call a *recovery*, happens when an existing neighbor dies (which is considered to be abrupt) and the successor of the failed neighbor takes over that zone to become the new neighbor of v.

We next define several additional metrics to facilitate explanation in later parts of the paper. Notice that one cycle in the life of a particular neighbor pointer is composed of several switches and one recovery as shown in Fig. 2(a). In the figure, thick horizonal lines represent online presence of peers that own v's neighbor pointer in the DHT space. The topmost line is the original neighbor with residual lifetime Z_1 acquired by v during join. As peers split the zone of the current neighbor, the link switches to two additional users. Once the last user dies at time R_1, the link is considered dead and a replacement process is initiated.[3] Recovery is complete after S time units when another node takes over the zone of the dead peer and is selected as v's new neighbor.

In all other aspects, the second recovery cycle behaves identical to the first one and leads to link failure after R_2 time units. This ON/OFF nature of a link process is shown in Fig. 2(b) where we assume that all repair delays S are $i.i.d.$ random variables, but the distribution of link lifetimes $R_1, R_2, \ldots$ may depend on the cycle number (in fact they do in certain cases studied below).

The final note is that it is important to distinguish the residual lifetime of the first neighbor from that of a link. While in non-switching systems the former metric (e.g., variables $Z_1, Z_2, \ldots$) determine how long a link stays alive, this is no longer the case in switching networks. Instead, the latter metric formalized as $R_1, R_2, \ldots$ determines query performance and a user's ability to tolerate churn. Our next step is to understand the behavior of these random variables under general lifetime distributions.

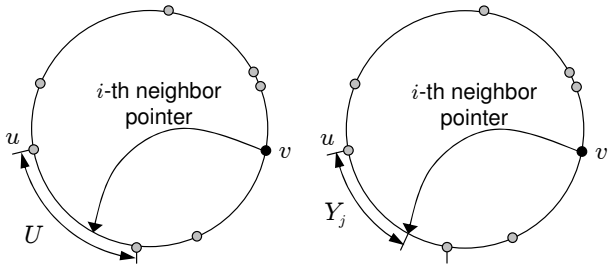
IV. Link Lifetime Model

In this section, we construct a semi-Markov model for the distribution of lifetimes $R_1, R_2, \ldots$ of a given link in a user's routing table.

A. Preliminaries

Recall that arriving users split zones of existing nodes based on a uniformly random hashing function. Denote by U the random zone size of existing users in a stationary system as shown in Fig. 3(a). Further assume that during join or the current recovery step that starts cycle j, successor u takes over pointer i as shown in Fig. 3(b). Then, define Y_j to be the

[3] Specifics of detecting failure are not essential to our results as repair delay is not studied in this paper.



(a) zone size U of user u (b) remaining zone size Y_j

Fig. 3. Zone size U and remaining zone size Y_j of user u.

remaining zone size between this pointer and the index of u. Intuitively, if the remaining zone Y_j is large, then it is likely that a new arrival will soon split the zone and the ownership of the link will be transferred to another peer. Therefore, link lifetimes are determined not by the distribution of U, but rather by that of Y_j. We derive both metrics later in the paper and next show how they can be used to obtain $R_1, R_2, \ldots$

For simplicity of notation, define *conditional link lifetime* $R(y)$ as the duration of the link conditioned on the fact that the remaining zone size Y_j is $y > 0$. Then, observe that the CDF of link lifetimes R_j can be written as:

$$P(R_j < x) = \int_0^\infty P(R(y) < x) f_{Y_j}(y) dy, \tag{1}$$

where $f_{Y_j}(y)$ is the probability density function (PDF) of remaining zone size Y_j (note that the distribution of Y_j depends on cycle number j). Similarly, we can obtain the expectation of R_j as:

$$E[R_j] = \int_0^\infty E[R(y)] f_{Y_j}(y) dy. \tag{2}$$

Thus, the task of deriving link lifetime R_j is reduced to analyzing the properties of conditional link lifetime $R(y)$ and the distribution of remaining zone size Y_j. In the rest of this section, we construct a semi-Markov process for each $R(y)$ and leave the derivation of the distribution of Y_j for deterministic DHTs to Section V and that for randomized DHTs to Section VI.

B. Conditional Link Lifetimes

For each zone size y, let variable A_δ^y count the number of switches (i.e., replacements by new users) that have occurred along the link in the time interval $[0, \delta]$, where time 0 denotes the instance when user v finds the first neighbor at the beginning of the current cycle. Denote by $A_\delta^y = F$ a special absorbing state into which A_δ^y arrives if the current neighbor attached to the link is in the failed state at time δ.

Then, it is easy to see that $\{A_\delta^y; \delta \geq 0\}$ is a continuous-time stochastic process with state space $\{F, 0, 1, 2, \ldots\}$ whose state transitions are shown in Fig. 4. As depicted in this figure, for each state $i \geq 0$, the process can jump into either state $i+1$, which means that a given zone is further split by a new arrival (i.e., the number of switches increases by 1), or state F, which

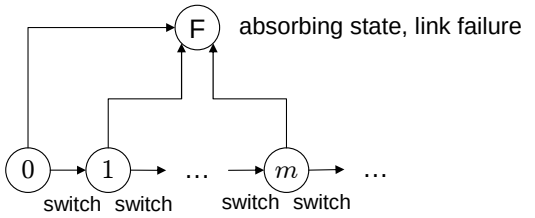


Fig. 4. State diagram for the process $\{A_\delta^j, \delta \geq 0\}$ of neighbor changes.

represents link failure. The initial state of the process at time 0 is always 0.

Using notation $\{A_\delta^y\}$, variable $R(y)$ can be described as the first-hitting time of process $\{A_\delta^y\}$ onto state F given that $A_0^y = 0$:

$$R(y) = \inf\{\delta > 0 : A_\delta^y = F | A_0^y = 0, Y_j = y\}. \tag{3}$$

The next theorem shows that $\{A_\delta^y; \delta \geq 0\}$ is a semi-Markov chain that describes the process of new users entering a given zone of initial length y and repeatedly splitting it.

Theorem 1: Process $\{A_\delta^y, \delta \geq 0\}$ for a given remaining zone size $Y_j = y$ is a regular semi-Markov chain. The sojourn time τ_i in state i follows the following general distribution:

$$P(\tau_i > x) = \begin{cases} P(W_0 > x) P(Z_j > x) & i = 0 \\ P(W_i > x) P(L > x) & i \geq 1 \end{cases}, \tag{4}$$

where Z_j is the residual lifetime of the first neighbor that starts the j-th cycle, L is user lifetime with CDF $F(x)$, W_i is an exponential random variable with rate λ_i:

$$\lambda_i = \frac{E[N] y}{E[L] 2^i}, \quad i \geq 0, \tag{5}$$

and $E[N]$ is the mean system size. Furthermore, transition probability $p_{i,i+1}$ from state i to $i+1$ is given by:

$$p_{i,i+1} = \begin{cases} P(W_0 < Z_j) & i = 0 \\ P(W_i < L) & i \geq 1 \end{cases}, \tag{6}$$

and the probability $p_{i,F}$ to absorb from state i is equal to $1 - p_{i,i+1}$.

Note from (5) that as the number of switches within a zone (i.e., variable i) increases, arrival rate λ_i into state i decreases exponentially fast and the mean waiting time W_i until the next arrival increases at the same rate. As state $i \to \infty$, $W_i \to \infty$ and thus process $\{A_\delta^y\}$ jumps into failed state F with probability $p_{i,F}$ that converges to 1.

Next, we study the distribution and expectation of conditional link lifetime $R(y)$. Denote the CDF of the sojourn time τ_i in state i by:

$$G_i(t) = P(\tau_i < t). \tag{7}$$

Noting from (4) that τ_i of chain $\{A_\delta^y\}$ is independent of the next state, the matrix of the semi-Markov kernel $Q(t) = [q_{ik}(t)]$ is given by [6]:

$$q_{ik}(t) = p_{ik} G_i(t), \quad i, k \in \{F, 0, 1, \ldots\}, \tag{8}$$

where p_{ik} is the transition probability from state i to state k given in (6). Taking the Laplace (Stieltjes) transform of $q_{ik}(t)$ leads to:

$$\hat{q}_{ik}(s) = \int_0^\infty e^{-st} dq_{ik}(t) = p_{ik} \int_0^\infty e^{-st} dG_i(t). \quad (9)$$

Define the following Laplace transform of the first hitting time $R(y)$ from state 0 to F as:

$$\hat{R}(s, y) = E[e^{-sR(y)}]. \quad (10)$$

Though it is known that the Laplace transform of the first-hitting time of a semi-Markov chain can be computed using spectral properties of kernel $Q(t)$ [3], this approach hides the effect of system parameters on the resulting distribution. Due to the simplicity of state transitions of chain $\{A_\delta^y\}$, we next derive $\hat{R}(s, y)$ without involving matrix operations on $Q(t)$.

Theorem 2: The Laplace transform $\hat{R}(s, y)$ of conditional link lifetime $R(y)$ is given by:

$$\hat{R}(s, y) = \hat{q}_{0F}(s) + \sum_{k=1}^{\infty} \left(\prod_{i=0}^{k-1} \hat{q}_{i,i+1}(s) \right) \hat{q}_{kF}(s), \quad (11)$$

where $\hat{q}_{ik}(s)$ are shown in (9).

With $\hat{R}(s, y)$ in hand, we can apply the inverse Laplace transform to retrieve the distribution of $R(y)$ and take the derivatives of $\hat{R}(s, y)$ to get its moments. Next, we use a simpler approach to obtain the mean $E[R(y)]$.

Theorem 3: The expected conditional link lifetime is:

$$E[R(y)] = E[\tau_0] + \sum_{k=1}^{\infty} \left(\prod_{i=0}^{k-1} p_{i,i+1} \right) E[\tau_k], \quad (12)$$

where $E[\tau_k]$ is the expected sojourn time in state k shown in (4) and $p_{i,i+1}$ are state transition probabilities in (6).

Theorems 1–2 demonstrate that variable $R(y)$ is fully determined by user lifetimes L and residual neighbor lifetimes Z_j. Our remaining steps are to analyze the properties of Z_j and derive the distribution of remaining zone sizes Y_j for both deterministic and randomized DHTs.

V. Deterministic DHTs

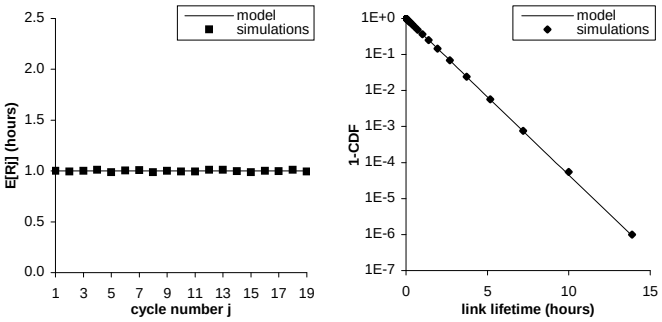
In deterministic DHTs, each neighbor pointer of user v is generated based on a fixed distance between the pointer and the user. We start this section by deriving a model for $R(y)$ under two types of user lifetimes and then analyze the distribution of residual zone size Y_j.

A. Residual Lifetimes of Neighbors

Using the user churn model summarized in Section III-A, it has been shown in [34, Theorem 3] that the equilibrium CDF $P(Z_1 < x)$ of residual neighbor lifetimes under random selection during join is a simple function of the user lifetime distribution $F(x)$:

$$P(Z_1 < x) = \frac{1}{E[L]} \int_0^x (1 - F(u)) du, \quad (13)$$

where $E[L]$ is the mean user lifetime. The next lemma shows that (13) also holds for all $j \geq 2$.



(a) $E[R_j]$ (b) Distribution of R_3

Fig. 5. Comparison of model (14) to simulations in a deterministic DHT with $E[N] = 2,000$ and exponential user lifetimes with $E[L] = 1$ hour.

Lemma 1: For all $j \geq 1$, the CDF of residual lifetime Z_j of the initial neighbor that starts the j-th cycle converges to (13) as system age approaches infinity.

It is important to emphasize that Lemma 1 holds when switching occurs in DHTs in response to *Poisson* user arrivals into the system and may not hold otherwise. When a neighbor pointer switches to a new user, it loses track of which peer on the ring will be the neighbor that will start the next cycle in the link's ON/OFF process. Hence, neighbor selection during link recovery is essentially uniformly random among the existing neighbors (due to random hash indexes) and independent of the selected neighbor's age.

Before we show simulation results in the next subsection, we define rules for generating a DHT under churn. In simulations, user arrivals follow a Poisson process with a constant rate $E[N]/E[L]$, where the mean system size $E[N]$ and the average user lifetime $E[L]$ are determined a-priori. Each user departs at the end of its lifetime L, which is drawn from a given distribution $F(x)$. In addition, each joining user obtains a uniformly random hash index in $[0, 1)$, follows the random-split algorithm during join, and performs recovery when its predecessors die. After the system has evolved for enough time, we compare simulation results to the derived models.

B. Conditional Link Lifetimes

We next use exponential and later Pareto lifetimes to study the behavior of $R(y)$. Assume that user lifetimes L are exponential with rate $\mu = 1/E[L]$. Then, it is easy to obtain from Lemma 1 that residual lifetime Z_j of the initial neighbor, for all cycles $j \geq 1$, is exponential with the same rate μ. Using $L \sim \exp(\mu)$ and $Z_j \sim \exp(\mu)$ and invoking Theorem 2 leads to the following result.

Theorem 4: For user lifetimes L with CDF $1 - e^{-\mu x}$, link lifetime R_j is independent of remaining zone size Y_j and has the same distribution as L:

$$P(R_j < x) = 1 - e^{-\mu x}, \quad \text{for all } j \geq 1, \quad (14)$$

where $\mu = 1/E[L]$.

Model (14) is very accurate as shown in Fig. 5. Notice from the left figure that $E[R_j]$ is equal to mean user lifetime $E[L]$ and from the right figure that the distribution of R_j is indeed

exponential, which holds for any $j \geq 1$ (only R_3 is shown in the figure).

The rationale behind Theorem 4 can be explained as follows. Recall that Z_j is the residual lifetime of the first neighbor u that owns the neighbor pointer in each cycle. Due to the memoryless property of exponential distributions, the remaining time of Z_j obtained at a random instant is still exponential with rate μ, which matches the lifetime distribution of new arrivals entering the same zone. Therefore, it makes no difference whether a current neighbor u is replaced by a new arrival or not. Then, it is not hard to see that the link lifetime has the same distribution as Z_j, which is $\exp(\mu)$. A similar scenario is observed in $M/M/1$ queues [33] where customers can be interrupted during services and the distribution of the total service time required for a customer does not change.

Theorem 4 indicates that switching has no impact on link lifetimes in any DHT with exponential user lifetimes, which makes analysis of system performance in such systems very simple. However, we should note that this result does not hold for any non-exponential lifetime distribution. As recent measurements of P2P networks show that user lifetimes are often heavy-tailed [4], [31], we next use the Pareto distribution $P(L < x) = 1 - (1 + x/\beta)^{-\alpha}$ with shape parameter $\alpha > 1$ and scale parameter $\beta > 0$ to estimate the performance of real DHTs under churn.

For Pareto lifetimes, it is clear from Lemma 1 that the residual lifetime Z_j of initial neighbors follows the CDF $P(Z_j < x) = 1-(1+x/\beta)^{-(\alpha-1)}$ for all $j \geq 1$, which shows that Z_j are also Pareto distributed but more heavy-tailed. Next, we apply Theorem 2 to obtain the Laplace transform $\hat{R}(y, s)$ and Theorem 3 to obtain the mean of $R(y)$.

Theorem 5: For Pareto lifetimes L, the mean conditional link lifetime $E[R(y)]$ is given by (12) with

$$E[\tau_i] = \beta e^{\lambda_i \beta} E_{\alpha_i}(\lambda_i \beta), \quad p_{i,i+1} = \lambda_i E[\tau_i] \tag{15}$$

where arrival rate λ_i is given in (5), $E_k(x) = \int_1^\infty e^{-xu}\, u^{-k} du$ is the generalized exponential integral, and

$$\alpha_i = \begin{cases} \alpha - 1, & i = 0 \\ \alpha & i \geq 1 \end{cases}. \tag{16}$$

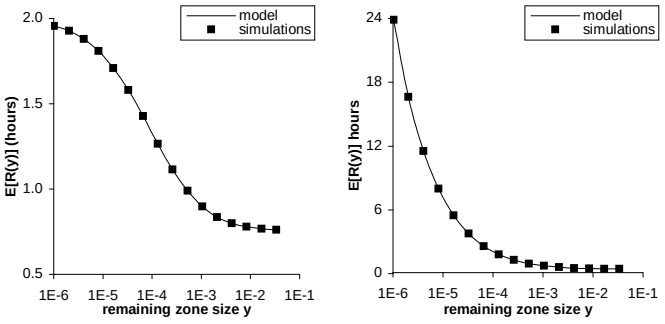
Furthermore, the Laplace transform $\hat{R}(y, s)$ is given by (11) with

$$\hat{q}_{i,i+1}(s) = \lambda_i E[\tau_i] A, \quad \hat{q}_{iF}(s) = (1 - \lambda_i E[\tau_i]) A, \tag{17}$$

where $A = 1 + (1 - \lambda_i - s)\beta e^{(\lambda_i+s)\beta} E_{\alpha_i}((\lambda_i + s)\beta)$, and $E[\tau_i]$ is shown in (15) and α_i in (16).

Fig. 6 shows that simulation results of $E[R(y)]$ for any fixed remaining zone size y are consistent with the model given in Theorem 5. Moreover, notice from this figure that as remaining zone size y reduces, $E[R(y)]$ increases and converges to $E[Z_1]$, where the distribution of neighbor residual lifetime Z_1 is given in (13).

We next derive the distribution of zone sizes in deterministic DHTs in order to obtain a computable model for R_j.



(a) $\alpha = 3$ (b) $\alpha = 1.5$

Fig. 6. Comparison of model $E[R(y)]$ in Theorem 5 to simulation results in a deterministic DHT with mean size $E[N] = 2,000$ and Pareto user lifetimes L with mean $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.

C. Zone Sizes

In order to determine the distribution of zone sizes U and Y_j in Fig. 3, we must decide on the zone splitting method. The derivations below only cover the random-split [32] mechanism (i.e., zones are split at hash indexes of arriving users) that is used in Chord [28] and only considers one-dimensional DHTs. A similar derivation can be carried out for the center-split [18], [24] strategy (i.e., zones are always split in the center) and multi-dimensional DHTs, but this analysis is much more tedious and is not shown here.

Since all arriving users are placed in the interval $[0, 1)$, the average zone size is approximately $1/E[N]$, where N is the random system size in the steady-state.[4] The next result states that in equilibrium DHTs, zone sizes no larger than $1/\sqrt{E[N]}$ are distributed approximately exponentially. Since most zone sizes do not deviate from the mean very far, this result directly applies to random variable U defined earlier.

Lemma 2: As the mean system size tends to infinity, the distribution of small zones in the DHT becomes approximately exponential:

$$\lim_{E[N]\to\infty} \frac{P(U > x)}{e^{-E[N]x}} = 1 \tag{18}$$

for all x such that $x^2 E[N] \to 0$.

Our next task is to obtain the distribution of remaining zone size Y_j in each cycle $j \geq 1$.

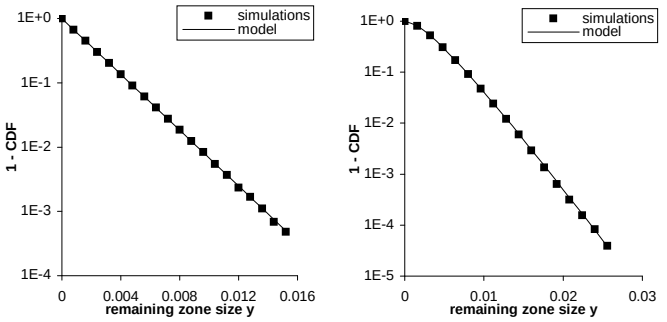
Lemma 3: For a given zone size y, assume that $y^2 E[N] \to 0$ as $E[N] \to \infty$. Then, the PDF $f_{Y_j}(y)$ of remaining zone size Y_j is asymptotically:

$$\begin{cases} \lim\limits_{E[N]\to\infty} \frac{f_{Y_1}(y)}{E[N]e^{-E[N]y}} = 1 & j = 1 \\ \lim\limits_{E[N]\to\infty} \frac{f_{Y_j}(y)}{E[N]^2 y e^{-E[N]y}} = 1 & j \geq 2 \end{cases}, \tag{19}$$

where $E[N]$ is the mean system size in equilibrium.

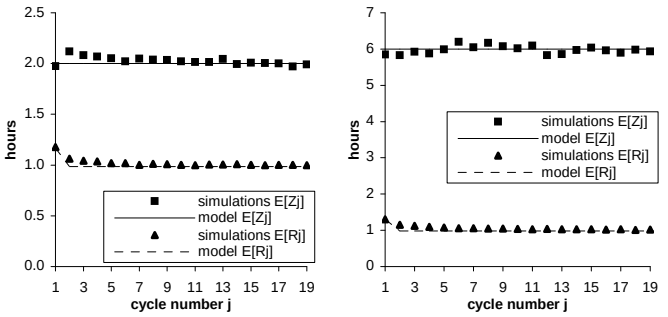
Lemma 3 shows that the distribution of Y_1 is exponential and that of Y_j for $j \geq 2$ is Erlang-2. As shown in Fig. 7, model (19) is very accurate even for small average system size

[4]Approximation $E[1/N] \approx 1/E[N]$ is asymptotically accurate as system size tends to infinity.



(a) Distribution of Y_1 (b) Distribution of Y_2

Fig. 7. Comparison of simulation results on the distribution of Y_j to model (19) in a deterministic DHT with mean size $E[N] = 500$ under churn introduced by Pareto L with $\alpha = 3$ and $E[L] = 1$ hour.



(a) $\alpha = 3$ (b) $\alpha = 2.2$

Fig. 8. Comparison of $E[R_j]$ to $E[Z_j]$ in a deterministic DHT with mean size $E[N] = 2,500$ users, Pareto lifetimes with mean $E[L] = 1$ hour, and $\beta = E[L](\alpha - 1)$.

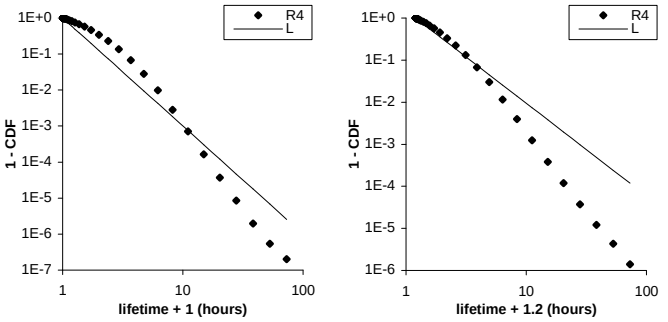
$E[N] = 500$ users. Additional simulation results confirming (19) for larger $E[N]$ and different j are not shown for brevity.

D. Putting the Pieces Together

The final step is to apply (1) and (2) to uncondition the distribution of link lifetime R_j and its mean $E[R_j]$ using the distribution of initial zone size Y_j given in (19). To this end, substituting $E[R(y)]$ shown in Theorem 5 and the PDF of Y_j in (19) into (2) leads to the final result on the mean link lifetime $E[R_j]$. Similarly, to get the distribution of R_j, we first retrieve the distribution of $R(y)$ from $\hat{R}(s, y)$ in Theorem 5 by applying an existing inverse Laplace transform software package [1]. Then substituting the distribution of $R(y)$ and (19) into (1) leads to the final model of the distribution of link lifetime R_j.

Fig. 8 shows simulations results and the model of the mean link lifetime $E[R_j]$ and the average residual lifetime $E[Z_j]$ of the initial neighbor that starts the j-th cycle. The model of $E[Z_j]$ is obtained using (13) and the general solution to $E[R_j]$ is given in (2). As shown in the figure, both models match simulation results very well and as α becomes smaller, the difference between $E[R_j]$ and $E[Z_j]$ increases as expected. The above results also show that the process of switching to new users can significantly reduce the lifetime of a link and that deterministic DHT systems with Pareto L can exhibit $E[R_j]$ very close to $E[L]$. This is in contrast to unstructured P2P systems where $E[R_j]$ can be $11 - 16$ times higher than $E[L]$ depending on shape parameter α [4], [31].

Further observe from the model and Fig. 8 that link lifetimes are completely characterized by two random variables R_1 and R_2 since R_j for $j \geq 3$ has the same distribution as R_2. This arises from the fact that zone size Y_1 is different from Y_2, while Y_j for $j \geq 3$ are all distributed as Y_2. Since Y_1 is stochastically smaller than Y_2 (see Lemma 3), it follows that R_1 is stochastically larger than R_2. Furthermore, from the analysis of the Markov chain in previous sections, it becomes clear that selecting neighbors with *smaller* initial zone sizes leads to larger link lifetimes since such neighbors are less likely to be replaced by newly arriving users and the link's $E[R_j]$ will be closer to $E[Z_j]$.



(a) $\alpha = 3$ (b) $\alpha = 2.2$

Fig. 9. Link lifetimes R_4 are less heavy-tailed than Pareto user lifetimes L in a deterministic DHT with mean size $E[N] = 2,500$ peers, $E[L] = 1$ hour, and $\beta = (\alpha - 1)E[L]$.

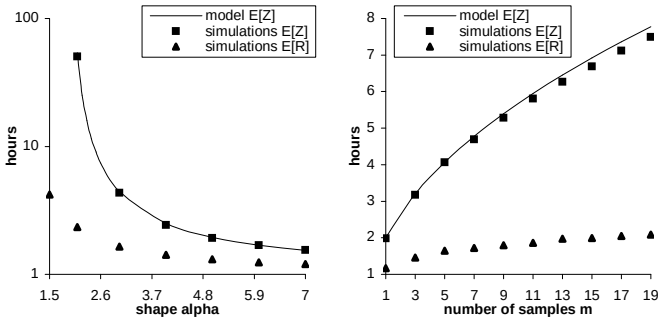
The most intriguing result shown in Fig. 8 is that $E[R_j]$ for all $j \geq 2$ is very close to the mean user lifetime $E[L]$ under different values of α (e.g., $E[R_4] = 0.986$ hours for $\alpha = 3$ and 1.096 for $\alpha = 2.2$). However, from the model of the tail distribution of link lifetime R_4 shown in Fig. 9, observe that the distribution of R_j for $j \geq 2$ is actually different from that of lifetime L and is *less* heavy-tailed than the original distribution. A similar result holds for other values of α and other distributions, which we do not show for brevity.

VI. Randomized Links

Our analysis suggests that since the user arrival process into a DHT is unchangeable, peers may utilize knowledge of residual lifetime Z_j of the initial owner of a given link and remaining zone size Y_j to improve link lifetime R_j. In the following, we make use of the freedom of selecting links in randomized DHTs to achieve the goal of increasing R_j using two link-selection strategies.

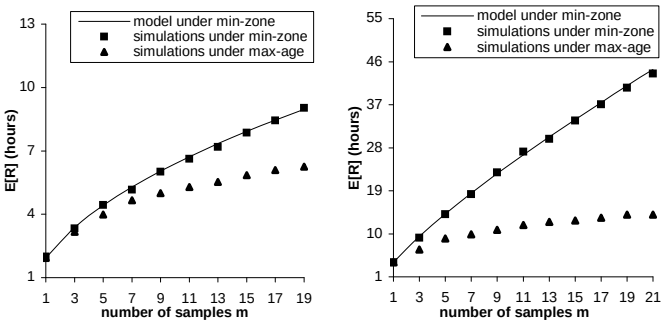
A. Max-Age Selection

The first strategy we apply for selecting neighbor pointers is called *max-age* [30], [35]. In this technique, which we explain using the example of Randomized Chord [10], user v with hash index $id(v) \in [0, 1)$ uniformly randomly samples m points in the range $[id(v) + 2^i/2^{64}, id(v) + 2^{i+1}/2^{64})$ and



(a) $m = 6$ (b) $\alpha = 3$

Fig. 10. Impact of shape α and number of samples m on mean link lifetime $E[R_j]$ under max-age selection in a randomized DHT with mean size $E[N] = 2,000$ for Pareto lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.



(a) $\alpha = 1.5$ (b) $\alpha = 1.2$

Fig. 11. Comparison of mean link lifetime $E[R_j]$ under min-zone selection to that under max-age selection in a randomized DHT with mean size $E[N] = 2,000$ for Pareto user lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.

selects the point whose successor has the maximum age as its i-th neighbor pointer. Note that switching occurs as described before (i.e., when new users split a given zone and replace existing neighbors) and link failure is repaired by replacing the dead neighbor (i.e., the last user holding the link) with the current successor.

It is clear that link lifetimes R_j for all cycles $j \geq 1$ have the same distribution since the neighbor pointer in each cycle is uniformly randomly generated within a certain range of users (as mentioned before, we assume the range is large enough to support non-trivial choices). Simulation results of max-age selection and the model of $E[Z_j]$ from [35] are shown in Fig. 10. First notice from part (a) that for a fixed number of samples $m = 6$, as shape α decreases, the mean link lifetime $E[R_j]$ increases much slower than the mean residual lifetime $E[Z_j]$ of the initial neighbor (in fact, $E[Z_j] = \infty$ for $\alpha \leq 2$). A similar phenomenon appears in part (b) where $E[Z_j]$ increases at the rate of $\sqrt{m}$ for $\alpha = 3$ (see [35, Lemma 5]), while $E[R_j]$ rises from 1.17 hours to only 2.09 hours as m increases from 1 to 19. These two subfigures demonstrate that the improvement in terms of the mean link lifetime $E[R_j]$ under max-age selection is generally very small since new arrivals sooner or later split initial neighbors to take ownership of the link and hence ages or residual lifetimes of original neighbors do not affect link churn rate very much.

B. Min-Zone Selection

To reduce the likelihood that new arrivals replace old neighbors when splitting a given zone, we propose a new strategy called *min-zone*. Similar to the max-age method, user v uniformly samples m points in $[id(v)+2^i/2^{64}, id(v)+2^{i+1}/2^{64})$, but then selects the point whose successor has the minimum zone size.

To obtain a model for $E[R_j]$ under min-zone selection, first note that residual lifetime Z_j of the initial neighbor starting the j-th cycle follows the distribution given in (13) since all m samples are uniformly random and zone sizes are independent of user ages or lifetimes. It is then clear that for a fixed remaining size $Y_j = y$, the Laplace transform and the mean conditional link lifetime given in Theorem 5 are both still valid. Next, given that initial zone size Y_j is minimum among m uniformly randomly selected samples, we readily obtain:

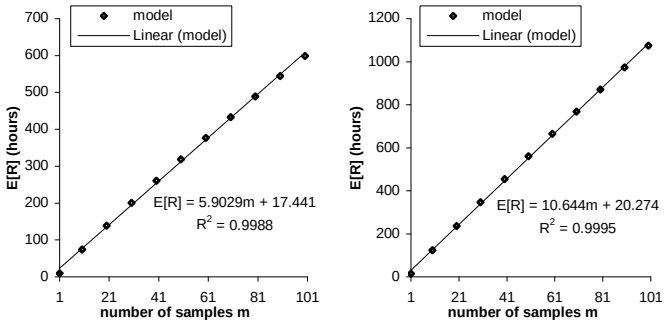
$$P(Y_j > y) = [P(U > y)]^m, \quad \text{for all } j \geq 1, \qquad (20)$$

where U is the zone size of a randomly selected user on the ring whose limiting distribution shown in (18). The final step is to combine Theorem 5 and (20) to obtain the distribution of R_j and its mean under min-zone selection.

As shown in Fig. 11, the model of $E[R_j]$ matches simulation results very well. Most interestingly, the figure demonstrates that the mean link lifetime $E[R_j]$ under min-zone selection is significantly larger than that under max-age selection for both choices of α and that the difference between the two metrics becomes more pronounced as the number of samples m increases or shape α decreases. Furthermore, this figure suggests that as $m \to \infty$, $E[R_j]$ for min-zone selection and $\alpha < 2$ goes to infinity, while $E[R_j]$ for max-age selection converges to some fixed number regardless of α. The following theorem confirms this result.

Theorem 6: For Pareto user lifetimes with $1 < \alpha \leq 2$, the expected link lifetime under min-zone selection approaches infinity for sufficiently large system population and random sample size: $\lim_{E[N]\to\infty} \lim_{m\to\infty} E[R_j] = \infty$. For max-age selection and any α, the mean link lifetime converges to a constant: $\lim_{E[N]\to\infty} \lim_{m\to\infty} E[R_j] < \infty$.

The above analysis indicates that min-zone selection is significantly better than max-age selection for very heavy-tailed user lifetimes. Since real systems have been observed to exhibit $\alpha \approx 1.06$ in [4] and $\alpha = 1.09$ in [31], this result paves a simple way for building better DHTs in practice. The amount of actual improvement in $E[R_j]$ for these two values of α is shown in Fig. 12, where the growth rate in both curves is approximately linear in m. The figures also show the corresponding linear fits to the model, which can be used to predict how m affects link lifetime $E[R_j]$ in these two cases. For instance, with $\alpha = 1.09$, users can obtain $E[R_j] \approx 76$ hours by sampling $m = 10$ points for each suitable (i.e., with enough random choices) link in a randomized DHT. For $\alpha = 1.06$, the corresponding average link lifetime is 127 hours.



(a) $\alpha = 1.09$ (b) $\alpha = 1.06$

Fig. 12. Approximation of $E[R_j]$ as a linear function of number of samples m under min-zone selection for Pareto user lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.

Comparing these numbers to $E[R_j] \approx E[L] = 1$ hour in deterministic DHTs, the extent of improvement is undoubtedly dramatic.

VII. Conclusion

This paper formalized the notion of "link lifetimes" in certain types of DHTs where link pointers switch to new neighbors in response to arriving peers. We introduced a semi-Markov process to model random replacement of neighbors along a given link and showed that lifetimes of deterministic links are much worse than those in unstructured P2P networks with heavy-tailed user lifetimes. For randomized DHTs, our results show that the proposed min-zone selection method is substantially more effective than the commonly-used max-age selection strategy and the mean link lifetime $E[R_j]$ under min-zone selection can be increased approximately linearly in the number of points m each user v samples.

Future work involves development of more sophisticated algorithms for increased DHT resilience, analysis of non-Poisson arrivals, and analysis of asymptotically small networks where limiting results similar to Theorem 6 do not hold.

References

[1] J. Abate and P. P. Valkó, "Multi-Precision Laplace Transform Inversion," *Int. J. Numer. Meth. Engng*, vol. 60, pp. 979–993, 2004.

[2] R. Bhagwan, S. Savage, and G. M. Voelker, "Understanding Availability," in *Proc. IPTPS*, Feb. 2003, pp. 256–267.

[3] J. T. Bradley, N. J. Dingle, P. G. Harrison, and W. J. Knottenbelt, "Distributed Computation of Passage Time Quantiles and Transient State Distributions in Large Semi-Markov Models," in *Proc. IPDPS*, Apr. 2003.

[4] F. E. Bustamante and Y. Qiao, "Friendships that Last: Peer Lifespan and its Role in P2P Protocols," in *Proc. Intl. Workshop on Web Content Caching and Distribution*, Sep. 2003.

[5] M. Castro, M. Costa, and A. Rowstron, "Performance and Dependability of Structured Peer-to-Peer Overlays," in *Proc. DSN*, Jun. 2004.

[6] E. Çinlar, *Introduction to Stochastic Processes*. Prentice Hall, 1997.

[7] Gnutella. [Online]. Available: http://www.gnutella.com/.

[8] P. B. Godfrey, S. Shenker, and I. Stoica, "Minimizing Churn in Distributed Systems," in *Proc. ACM SIGCOMM*, Sep. 2006.

[9] P. B. Godfrey, Personal Communication, 2006.

[10] K. Gummadi, R. Gummadi, S. Gribble, S. Ratnasamy, S. Shenker, and I. Stoica, "The Impact of DHT Routing Geometry on Resilience and Proximity," in *Proc. ACM SIGCOMM*, Aug. 2003, pp. 381–394.

[11] S. Krishnamurthy, S. El-Ansary, E. Aurell, and S. Haridi, "A Statistical Theory of Chord under Churn," in *Proc. IPTPS*, Feb. 2005, pp. 93–103.

[12] D. Leonard, V. Rai, and D. Loguinov, "On Lifetime-Based Node Failure and Stochastic Resilience of Decentralized Peer-to-Peer Networks," in *Proc. ACM SIGMETRICS*, Jun. 2005, pp. 26–37.

[13] D. Leonard, Z. Yao, X. Wang, and D. Loguinov, "On Static and Dynamic Partitioning Behavior of Large-Scale Networks," in *Proc. IEEE ICNP*, Nov. 2005, pp. 345–357.

[14] J. Li, J. Stribling, T. M. Gil, R. Morris, and M. F. Kaashoek, "Comparing the Performance of Distributed Hash Tables under Churn," in *Proc. IPTPS*, Feb. 2004, pp. 87–99.

[15] J. Li, J. Stribling, R. Morris, and M. F. Kaashoek, "Bandwidth-Efficient Management of DHT Routing Tables," in *Proc. USENIX NSDI*, May 2005, pp. 1–11.

[16] J. Li, J. Stribling, R. Morris, M. F. Kaashoek, and T. M. Gil, "A Performance vs. Cost Framework for Evaluating DHT Design Tradeoffs under Churn," in *Proc. IEEE INFOCOM*, Mar. 2005, pp. 225–236.

[17] D. Liben-Nowell, H. Balakrishnan, and D. Karger, "Analysis of the Evolution of the Peer-to-Peer Systems," in *Proc. ACM PODC*, Jul. 2002, pp. 233–242.

[18] D. Loguinov, A. Kumar, V. Rai, and S. Ganesh, "Graph-Theoretic Analysis of Structured Peer-to-Peer Systems: Routing Distances and Fault Resilience," in *Proc. ACM SIGCOMM*, Aug. 2003, pp. 395–406.

[19] G. Manku, M. Bawa, and P. Raghavan, "Symphony: Distributed Hashing in a Small World," in *Proc. USITS*, Mar. 2003, pp. 127–140.

[20] G. S. Manku, M. Naor, and U. Weider, "Know thy Neighbor's Neighbor: the Power of Lookahead in Randomized P2P Networks," in *Proc. ACM STOC*, Jun. 2004, pp. 54–63.

[21] P. Maymounkov and D. Mazieres, "Kademlia: A Peer-to-Peer Information System Based on the XOR Metric," in *Proc. IPTPS*, Mar. 2002, pp. 53–65.

[22] M. Naor and U. Wieder, "Novel Architectures for P2P Applications: the Continuous-Discrete Approach," in *Proc. ACM SPAA*, Jun. 2003, pp. 50–59.

[23] G. Pandurangan, P. Raghavan, and E. Upfal, "Building Low-Diameter Peer-to-Peer Networks," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 6, pp. 995–1002, Aug. 2003.

[24] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A Scalable Content-Addressable Network," in *Proc. ACM SIGCOMM*, Aug. 2001, pp. 161–172.

[25] S. Rhea, D. Geels, T. Roscoe, and J. Kubiatowicz, "Handling Churn in a DHT," in *Proc. USENIX Ann. Tech. Conf.*, Jun. 2004, pp. 127–140.

[26] A. Rowstron and P. Druschel, "Pastry: Scalable, Decentralized Object Location and Routing for Large-Scale Peer-to-Peer Systems," in *Proc. IFIP/ACM International Conference on Distributed Systems Platforms (Middleware)*, Nov. 2001, pp. 329–350.

[27] S. Saroiu, P. K. Gummadi, and S. D. Gribble, "A Measurement Study of Peer-to-Peer File Sharing Systems," in *Proc. SPIE/ACM Multimedia Computing and Networking*, vol. 4673, Jan. 2002, pp. 156–170.

[28] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications," in *Proc. ACM SIGCOMM*, Aug. 2001, pp. 149–160.

[29] D. Stutzbach and R. Rejaie, "Understanding Churn in Peer-to-Peer Networks," in *Proc. ACM IMC*, Oct. 2006, pp. 189–202.

[30] G. Tan and S. Jarvis, "Stochastic Analysis and Improvement of the Reliability of DHT-based Multicast," in *Proc. IEEE INFOCOM*, May 2007, pp. 2198–2206.

[31] X. Wang, Z. Yao, and D. Loguinov, "Residual-Based Measurement of Peer and Link Lifetimes in Gnutella Networks," in *Proc. IEEE INFOCOM*, May 2007, pp. 391–399.

[32] X. Wang, Y. Zhang, X. Li, and D. Loguinov, "On Zone-Balancing of Peer-to-Peer Networks: Analysis of Random Node Join," in *Proc. ACM SIGMETRICS*, Jun. 2004, pp. 211–222.

[33] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Prentice Hall, 1989.

[34] Z. Yao, D. Leonard, X. Wang, and D. Loguinov, "Modeling Heterogeneous User Churn and Local Resilience of Unstructured P2P Networks," in *Proc. IEEE ICNP*, Nov. 2006, pp. 32–41.

[35] Z. Yao, X. Wang, D. Leonard, and D. Loguinov, "On Node Isolation under Churn in Unstructured P2P Networks with Heavy-Tailed Lifetimes," in *Proc. IEEE INFOCOM*, May 2007, pp. 2126–2134.