Understanding Disconnection and Stabilization of Chord

Zhongmei Yao, *Member*, *IEEE*, and Dmitri Loguinov, *Senior Member*, *IEEE*

Abstract—Previous analytical work [16], [17] on the resilience of P2P networks has been restricted to disconnection arising from simultaneous failure of all neighbors in routing tables of participating users. In this paper, we focus on a different technique for maintaining consistent graphs—Chord's successor sets and periodic stabilizations—under both static and dynamic node failure. We derive closed-form models for the probability that Chord remains connected under both types of node failure and show the effect of using different stabilization interval lengths (i.e., exponential, uniform, and constant) on the probability of partitioning in Chord.

Index Terms—Peer-to-peer networks, graph disconnection, stabilization of Chord.

1 Introduction

Peer-to-peer (P2P) networks have received tremendous interest in recent years among both Internet users and computer networking professionals. One of fundamental problems in the study of these systems is the ability of the network to stay connected under node failure [1], [3], [7], [9], [10], [11], [15], [16], [20], [22], [24], [26], [29], [33]. While previous analytical work [16], [17] on disconnection of P2P networks has focused on neighbor tables and partitioning arising from failure of entire routing tables, structured P2P networks usually maintain auxiliary sets called *successor lists* [25], [26], whose sole purpose is to recover the system from inconsistent states and provide resilience [26]. In this paper, we focus on the partitioning of a widely used Distributed Hash Table (DHT) called Chord [26] whose successor rules are easily susceptible to analytical modeling. Note that similar results can be obtained for other types of successor/leaf sets.

Recall that Chord [26] maps users (and keys of data items) using a uniform hashing function onto the Chord ring $\{0, 1, \ldots, 2^m - 1\}$, where m is some sufficiently large number that can accommodate all nodes without conflict. Each user v maintains a *successor list* and a *routing table*. Assuming n peers in the system, the former set contains $r = \Theta(\log n)$ peers immediately following v along the ring and the latter set consists of $k = \Theta(\log n)$ neighbor pointers, where the ith neighbor is the first node following the point $id(v) + 2^{i-1}$ on the ring, $id(v)$ is the hash index of v, and $i = 1, 2, \ldots, k$. Note that routing tables are used to reduce lookup latency, while successors ensure resilience in the face of node failure. Even if all routing tables are in the failed state, Chord is still able to function by forwarding queries, repairing failures, and finding new neighbors via successor lists. In contrast, when all r successors of any node fail simultaneously, the system becomes partitioned and is potentially unable to recover without a bootstrap.[1] We generally call the event of a user losing all of its successors *node isolation* and note that it determines the likelihood of graph partitioning:

$$P(\text{graph disconnects}) = \mathrm{P}(\mathrm{X} > 0), \tag{1}$$

where X is the number of users that are isolated in the system. Due to the strong dependency among successor lists of consecutive users along the circle and entirely different stabilization strategies studied in this paper, previous neighbor churn models [16] cannot be applied to obtain the probability in (1). We perform this task below for both static and dynamic node failure.

1.1 Static Failure

Many prior studies have been interested in the resilience of structured P2P networks against *static* node failure [10], [11], [26], i.e., when each node independently fails with a certain probability p. Applying the Erdös-Rényi theorem [4], we show that under p-fraction node failure, the probability that Chord with size $n \to \infty$ remains connected is asymptotically:

$$\lim_{n\to\infty} \frac{P(X=0)}{e^{-n(1-p)p^r}} = 1, \tag{2}$$

where $r = \Theta(\log n)$ is the number of immediate successors a user monitors. It is rather surprising to find from (2) that although the dependency among successor lists of consecutive users is very strong, Chord enjoys the same level of static resilience as networks where connectivity is determined using routing tables consisting of largely independent neighbors [17]. We further show that the rationale behind this can be explained using the Chen-Stein method [2].

Setting $r = c \log_2 n$, where $c > 0$ is a constant, (2) indicates that as $n \to \infty$, the probability that Chord remains connected approaches 1 if $c > -1/\log_2 p$ and 0 if $c < -1/\log_2 p$. Note that this holds for both complete and incomplete Chord graphs.

1. Although neighbors in some routing tables may still be alive, there is no guarantee that the system can return to a consistent state after partitioning.

- *Z. Yao is with the Department of Computer Science, The University of Dayton, 300 College Park, Dayton, OH 45469-2160. E-mail: zyao@udayton.edu.*
- *D. Loguinov is with the Department of Computer Science, Texas A&M University, TAMU 3112, College Station, TX 77843-3112. E-mail: dmitri@cs.tamu.edu.*

Manuscript received 12 Oct. 2009; revised 2 Feb. 2010; accepted 7 Mar. 2010; published online 27 May 2010.
Recommended for acceptance by Y. Liu.
For information on obtaining reprints of this article, please send e-mail to: tpds@computer.org, and reference IEEECS Log Number TPDS-2009-10-0507.
Digital Object Identifier no. 10.1109/TPDS.2010.114.

1.2 Dynamic Failure

As observed in deployed structured P2P file-sharing systems [23], [27], users continually join and fail at a high rate of churn. The second part of this paper focuses on the connectivity of Chord under dynamic node failure. We assume that each joining user v obtains r clockwise closest peers as its successor list, and then, stays in the system for L time units, where L is drawn from some user lifetime distribution $F(x)$. To handle abrupt node departures (i.e., failure), v stabilizes its successor list every S time units, where S can be random or constant, and brings the number of successors back to r after each stabilization. For a particular stabilization to be successful, at least one user among r successors must stay alive for the entire interval S.

Assuming exponential user lifetimes L and exponential intervals S, we show that probability ϕ that node v is isolated due to simultaneous failure of its r successors within v's lifetime is upper bounded by

$$\phi \leq \frac{\rho\rho! r!}{(\rho + r)!}, \tag{3}$$

where $\rho = E[L]/E[S]$. Furthermore, we prove that as $\rho \to \infty$ (e.g., $E[S]$ becomes sufficiently small), the above upper bound becomes exact.

We then examine how individual node isolations affect partitioning of the system as nodes continuously join and leave. Using the Chen-Stein method, we establish that when $r \to \infty$, the probability that Chord stays connected after experiencing N user joins is asymptotically:

$$\lim_{N \to \infty} \frac{P(X = 0)}{(1 - \phi)^N} = 1, \tag{4}$$

where ϕ is the node isolation probability given in (3). This result shows that isolations of individual users in Chord can be treated as *independent* when system size and successor lists become large. While a similar phenomenon has been observed in [17] without proof for independent neighbor behavior in routing tables, our result in (4) is again for dependent node isolations and is formally proven.

As (4) indicates that the task of studying global connectivity can be reduced to that of local connectivity, we next focus on isolation probability ϕ under different stabilization strategies. We derive closed-form models of ϕ for uniform and constant S, both of which have been suggested for use in Chord [26]. Our results show that both stabilization strategies are much better than the exponential S suggested in [14], often reducing ϕ by several orders of magnitude. We further show that constant stabilization delays S are *optimal* and keep Chord's isolation probability as $E[S] \to 0$ approximately equal to:

$$\phi \approx \frac{\rho\rho!}{(\rho + r)!}, \tag{5}$$

where $\rho = E[L]/E[S]$. The amount of improvement over the exponential version (3) of this metric is by a factor of $r!$, which is significant in most cases.

We finish the paper by studying nonexponential lifetimes observed in real P2P graphs [30]. Even though models of ϕ for heavy-tailed (e.g., Pareto) user lifetimes are currently intractable, we show that ϕ in such systems is upper bounded by the exponential metric (3) and demonstrate the distance to the upper bound in simulations. We further show that the conclusion that stabilization intervals with smaller variance lead to smaller ϕ holds for Pareto lifetimes as well.

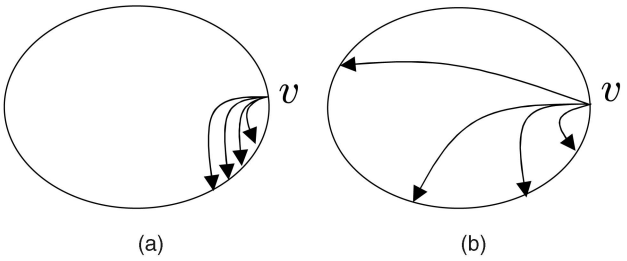

Fig. 1. User v's successors and neighbors in Chord. (a) r successors. (b) k finger table links.

The rest of the paper is organized as follows: Section 2 provides the basics of Chord and related work on static and dynamic node failure. In Section 3, we derive closed-form results on static resilience of Chord. In Section 4, we formalize the successor list model and derive node isolation probability and graph disconnection probability in Chord under dynamic node failure. Section 5 deals with node isolation probability under different stabilization strategies and finds the optimal method to keep Chord connected with the highest probability. In Section 6, we show simulation results on isolation probability for heavy-tailed user lifetimes. Section 7 concludes this paper.

2 Background

2.1 Chord Basics

In Chord, each key is assigned to the *successor* node, i.e., the first peer whose identifier is larger than the key in the clockwise direction along the ring. As illustrated in Fig. 1, each node v's *successor list* consists of the first r consecutive nodes following v, while finger pointers in v's routing table connect to neighbors with exponentially increasing distances to v.

Finger tables are used during key lookup where the originating node performs jumps of exponentially decreasing length until it finds the node responsible for the key or encounters an inconsistent state (e.g., stale pointer and dead successor) at one of the intermediate nodes. Inconsistent states in finger tables and successor lists are periodically repaired using a *stabilization technique*, which allows Chord to fix links broken during user departure, detect new peer arrival, and ensure lookup success during churn. When any node v leaves the system, its predecessor u notices v's departure during its periodic stabilization. Peer u then replaces v with the next alive user along the circle and adjusts its successor list accordingly. This process tolerates multiple nodes failing simultaneously and only requires that no successor list sustains a failure of all r nodes within a given stabilization interval. Similarly, node v learns of new arrivals during its stabilization process and properly adjusts its successor list to include the new peers.

Successor lists are generally used in routing only during the last step of a lookup or when all finger pointers corresponding to desired jump lengths have failed. As long as each node has at least one live peer in its successor list, the system is able to correct (after some delay) all stale finger pointers and repopulate each successor list with r correct entries, thus ensuring consistency and efficiency of subsequent lookups. However, when the entire successor list of any user v fails, Chord becomes *partitioned* [26]. The goal of this paper is to understand disconnection of Chord in the face of node failure and find the best stabilization algorithm that keeps Chord connected with the highest probability.

2.2 Resilience under Node Failure

Performance of DHTs under p-fraction node failure (i.e., static node failure) [10], [11], [26] and churn (i.e., dynamic node failure) [6], [14], [18], [19], [20], [21], [24] have received significant attention since the advent of structured P2P networks. While the problem of connectivity under failure for general graphs remains NP-complete [8], [12], [28], recent work [17] shows that several types of deterministic and random networks remain connected if and only if they do not develop isolated nodes after the failure. Despite its importance, the methodology in [17] only considers the resilience of *neighbor tables* rather than that of successors and does not model stabilization. The issues studied in this paper are analytically different due to the much stronger dependency between successor lists of neighboring nodes than between their finger tables and the fact that stabilization requires an entirely different model than the one in [17].

Another modeling work by Krishnamurthy et al. [14] studies the probability of finding a neighbor or successor in one of its three states (alive, failed or incorrect) and uses this model to predict lookup consistency and latency for exponential user lifetimes and exponential stabilization intervals.

3 Static Node Failure

In this section, we tackle resilience of Chord under *static node failure*, which means that the system sustains a one-time simultaneous failure event where each user becomes dead with an independent probability p. This analysis introduces a new model of handling-dependent random events in Chord and can be applied to systems of nonhuman entities (e.g., file systems) where failures can, in fact, be synchronized. The next section covers the more typical case of user churn observed in human-based P2P systems.

3.1 Basic Asymptotic Model

Suppose that Chord is in a consistent state such that each node correctly links to its r closest successors. Under static node failure, p fraction of nodes in the system fail simultaneously, where $0 \le p \le 1$ is a given number [10], [11], [17], [26]. Define a Bernoulli random variable X_i indicating whether node i is *isolated* due to the fact that its r successors all fail while i survives:

$$X_i = \begin{cases} 1, & \text{user } i \text{ is alive and its } r \text{ successors failed,} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Unlike [17], our definition does not involve finger tables since we are only interested in disconnection/isolation arising from disrupted successor lists. Then, the number of isolated nodes X in the system is the sum of a large number of *dependent* random variables X_i:

$$X = \sum_{i=1}^{n} X_i, \tag{7}$$

where n is the number of nodes in Chord. It is then clear from (1) that the probability that Chord remains connected (i.e., is not partitioned) is equal to $P(X = 0)$. The next theorem provides an asymptotic closed-form expression of $P(X = 0)$; however, we should note that this result is very different from similar analysis in [17] for two reasons: 1) the model in [17] only considers variables X_i with diminishing dependency as $r \to \infty$, which is not the case here; 2) the final result on the behavior of X is given in [17] without a formal proof due to a much wider variety of neighbor sets covered by [17].

Theorem 1. *The probability that each user in Chord remains connected to at least one successor under p-fraction node failure is asymptotically:*

$$\lim_{n\to\infty} \frac{P(X=0)}{e^{-n(1-p)p^r}} = 1, \tag{8}$$

where r is the number of successors at each node.

Proof. Denote by a Bernoulli random variable Y_i the event that node i has failed. Then, we have

$$p = P(Y_i = 1) = 1 - P(Y_i = 0). \tag{9}$$

Define L_n to be the length of the longest consecutive run of 1s in sequence $\{Y_1, \ldots, Y_n\}$:

$$L_n = \max_{1\le i\le n-k+1} \{k : Y_i = Y_{i+1} = \cdots = Y_{i+k-1} = 1\}. \tag{10}$$

Now, note that computing $P(X = 0)$ can be reduced to finding the distribution of L_n and ensuring that no run longer than $r - 1$ peers exists:

$$P(X = 0) = P(L_n < r). \tag{11}$$

Given that $r = \Theta(\log n)$ so that $r \to \infty$ as $n \to \infty$, the distribution of L_n converges to the following based on the Erdös and Rényi law [4]:

$$\frac{P(L_n < r)}{e^{-n(1-p)p^r}} \to 1, \tag{12}$$

as $n \to \infty$, which immediately leads to (8). □

We now make several notes about this result. First, observe that L_n defined in (10) can be heuristically interpreted as the maximum of a large number of geometric variables. Indeed, under static node failure, variables $\{Y_i\}$ in (9) are independent Bernoulli trials, e.g., $\{0, 1, 0, 0, 1, 1, \ldots\}$. Denote by K the random number of 0s (i.e., alive nodes) among n Bernoulli trials and by $W_1, W_2, \ldots, W_K$ the length of consecutive runs of 1s (i.e., node failures) on Chord ring. By the strong law of large numbers, as $n \to \infty$, $K/n \to 1 - p$ and $(W_1 + \cdots + W_K)/n \to p$ almost surely. Therefore, L_n is the maximum of *approximately* $n(1 - p)$ W_is:

TABLE 1
Comparison of Simulation Results of $P(X=0)$ under Static Node Failure to Model (15) in Chord

$p=.933$, $r=\lceil 2\log_2 n\rceil$			$n=50,000$, $r=\lceil 2\log_2 n\rceil$			$n=50,000$, $r=\lceil 10\log_2 n\rceil$			$n=50,000$, $r=\lceil\sqrt{n}\rceil$		
n	Simulations	(15)	p	Simulations	(15)	p	Simulations	(15)	p	Simulations	(15)
1,000	.9417	.9369	.5	1.0000	1.0000	.89	.9999	.9999	.92	1.0000	1.0000
5,000	.9373	.9360	.55	.9999	.9999	.9	.9997	.9997	.93	.9997	.9997
10,000	.9367	.9360	.6	.9983	.9984	.91	.9983	.9983	.94	.9971	.9971
20,000	.9365	.9360	.65	.9821	.9821	.92	.9919	.9918	.95	.9747	.9747
30,000	.9368	.9367	.70	.8472	.8473	.93	.9614	.9613	.96	.8077	.8076
40,000	.9363	.9361	.71	.7771	.7771	.94	.8344	.8343	.97	.1950	.1954
50,000	.9393	.9393	.75	.2850	.2849	.95	.4514	.4514	.98	.0000	.0000
100,000	.9395	.9394	.79	.0038	.0038	.96	.0368	.0371	.99	.0000	.0000

$$L_n \approx \max(W_1,\ldots,W_{n(1-p)}). \tag{13}$$

Recalling that the geometric CDF $P(W_i \le r-1) = 1-p^r$ if $r-1<np$, we then have that for $n\to\infty$,

$$P(L_n \le r-1) \approx (1-p^r)^{n(1-p)} \le e^{-n(1-p)p^r}. \tag{14}$$

Now, note that the upper bound in (14) becomes tight as $n(1-p)$ increases and/or p^r decreases. Our second note is that [4] provides an independent, but much more complex proof for the asymptotical result in (12). Finally, observe that (8) is valid for $r<np$ as long as $n(1-p)\to\infty$, which indicates that Theorem 1 holds for many choices of $r<np$, not only $r=\Theta(\log n)$ as in Chord.

The asymptotic result in (8) allows us to utilize a very accurate approximation:

$$P(\text{Chord is connected}) = P(X=0) \approx e^{-n(1-p)p^r}, \tag{15}$$

which we verify next in finite-size graphs. Simulation results of $P(X=0)$ in Chord under static node failure are presented in Table 1. In simulations, each node selects its node ID according to a uniform hashing function and connects to its r successors. After p fraction of users are uniformly randomly chosen and removed, the graph is checked to see how many users X are isolated. Notice from the first three columns in Table 1 that simulation results with $r=\lceil 2\log_2 n\rceil$ and $p=2^{-1/2}=0.993$ show that as n increases from 1,000 to 10,000, the discrepancy between model (15) and simulation results reduces fast. The rest of the table shows additional examples of model's accuracy for several choices of p and r.

It is interesting to see from (15) that if the length of a successor list r is independent of n, then the probability that the system remains connected approaches 0 as $n\to\infty$ (since in this case, $e^{-n(1-p)p^r}\to 0$). Therefore, r must grow when n increases in order to ensure the network connectivity under static node failure. Letting $r=c\log_2 n$ for Chord ring, where $c>0$ is a constant, we obtain that

$$\exp\bigl(-n(1-p)p^{c\log_2 n}\bigr) = \exp\bigl(-(1-p)n^{1+c\log_2 p}\bigr). \tag{16}$$

It is then easy to notice from (16) that $P(X=0)\to 0$ if $1+c\log_2 p>0$ and $P(X=0)\to 1$ if $1+c\log_2 p<0$. This demonstrates that as system size becomes large, condition $c>-1/\log_2 p$ is both necessary and sufficient for Chord to stay connected with probability close to 1.

3.2 Discussion

We next relate our results in Theorem 1 to those in [17, Proposition 3]. Recall that [17] defines isolation as an event of a user losing all of its neighbors in Fig. 1b. Their results show that all users have at least one alive neighbor with probability:

$$P(X=0) \approx e^{-n(1-p)p^k}, \tag{17}$$

where n is the system size, p is the independent node failure probability, and k is the number of neighbors in each node's table. Note that we have obtained an almost identical result (15) for successor lists in Chord, which is rather surprising since the dependency among isolation of nodes in Chord is much more significant than assumed in [17] (e.g., node i and node $i+1$ in Chord share $r-1$ common successors).

In fact, observe that the probability that node i is isolated due to the failures of its r successors is simply:

$$\phi = P(X_i=1) = (1-p)p^r, \quad 1\le i\le n, \tag{18}$$

where X_i is the Bernoulli variable defined in (6). Note that given that $r\to\infty$ as $n\to\infty$, it is readily seen from (18) that $\phi\to 0$ as $n\to\infty$. Using (18), the approximation in (15) can be transformed into:

$$P(X=0) \approx e^{-n\phi} \approx (1-\phi)^n, \tag{19}$$

where Taylor expansion $e^{-x}=1-x$ holds for small enough x as $n\to\infty$. Thus, (19) indicates that

$$P(X=0) = P\left(\bigcap_{i=1}^{n}[X_i=0]\right) \approx \prod_{i=1}^{n} P(X_i=0), \tag{20}$$

as $n\to\infty$, which shows that variables X_i in Chord behave *as if* they are completely independent. Note that when $r\to\infty$ as $n\to\infty$, node isolations become rare events. Then, (20) can be explained by the Chen-Stein theorem [2], which proves that the number of occurrences of dependent rare events X_i is approximately a Poisson random variable under certain conditions (this method will be explicitly used in the next section when we discuss these conditions). Therefore, as $n\to\infty$, Chord asymptotically exhibits the *same* static resilience using its successor lists composed of largely dependent users as other P2P networks using mostly independent peers in their neighbor sets [17]. However, the rate of convergence of $P(X=0)$ in (15) and (17) is different.

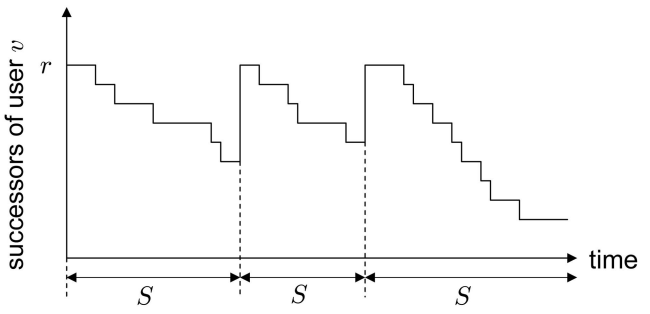

Fig. 2. Evolution of a node's successor list over time.

4 Dynamic Node Failure: General Results

Recent measurements of P2P networks [5], [23], [27], [30] show that peers continuously join and depart the system, which is often called *churn*. Thus, unlike static node failures which happen simultaneously, node failures in human-based P2P networks often occur dynamically as the system evolves over time. In this section, we first introduce the successor list model under churn, examine probability ϕ that all successors of node vs fail within its lifetime, and then derive the probability that Chord remains connected when stabilization intervals are exponentially distributed. We leave derivations for nonexponential intervals for the next section.

4.1 Successor List Model

When each user v joins the system, it acquires a successor list with r nearest nodes, and then, maintains it through periodic stabilizations (i.e., checks for consistency and dead users). We assume that v does not attempt to track failure of individual users as soon as they occur, but rather performs stabilization every S time units on the entire successor list (i.e., as done in Chord). At each stabilization interval, v corrects its successor list by skipping over failed nodes and appropriately adding to the list new arrivals (if any) [20], which always brings the number of successors at the end of stabilization back to r as long as the system has not been disconnected at some earlier time. For stabilization to be successful, at least one user among r successors must survive the entire stabilization interval. The interval S between two successive stabilizations reflects the duration needed to complete network-related activity to detect failure, exchange neighbor information, and any stabilization rate-limiting applied by the nodes.

Fig. 2 illustrates the evolution of user v's successor list in our simple model. As shown in this figure, the number of successors is r in the beginning of each stabilization interval of size S. This number then monotonically decreases over time until the next interval starts. If all r successors fail within any interval S before v departs, v is isolated and Chord is disconnected.

In general, as users continuously join and leave the system, the evolution of a node's successor list is rather complicated. It involves not only newly arriving users that replace existing successors, but also remaining lifetimes of existing successors at the start of each stabilization interval. For exponential lifetime distributions $F(x)$, however, user disconnection under this successor list model becomes tractable as we show next.

Before we proceed with derivations, we introduce the rules for running simulations whose purpose in this paper is to verify model accuracy under real-life conditions (i.e., P2P systems of *finite* age and size). In simulations, user arrivals occur according to a Poisson process. The rate of this arrival process is given by $E[N]/E[L]$, where $E[N]$ is the mean system size in equilibrium and $E[L]$ is the mean user lifetime. When a new user joins the system, it is assigned a uniformly random ID in the set $\{0, 1, \ldots, 2^{32} - 1\}$ and given r immediate successors. Each user then monitors its r successors, stabilizes them every S-interval, and departs from the system after L time units, where L is drawn from some user lifetime distribution $F(x)$. We use exponential $F(x) = P(L \le x) = 1 - e^{-x/E[L]}$ for the majority of closed-form results below, but later extend our analysis to Pareto $F(x) = 1 - (1 + x/\beta)^{-\alpha}$ [27], [31] in Section 6.

4.2 Node Isolation

Denote by $Z(t)$ the number of successors of node v at time t, where $t = 0$ is the time when v joins the system. Note that $Z(0) = r$ and $Z(t) \le r$ at any age t. In the following, we show that $\{Z(t)\}$ is a Markov chain for exponential user lifetimes and exponential stabilization intervals, which is followed by the derivation of the exact model of node isolation probability ϕ. This exact model is necessary for verifying the accuracy of our later closed-form bounds on ϕ.

Observe from Fig. 2 that state transitions of process $\{Z(t)\}$ are triggered by either failure of existing successors or stabilizations that occur at rate of $\theta = 1/E[S]$. Due to the memoryless property of exponential lifetime distributions, the failure rate of each existing successor (no matter old or new) is $\mu = 1/E[L]$, which is the key reason that makes the successor list tractable for exponential L. This leads to the following lemma:

Lemma 1. *For exponential lifetimes $L \sim \exp(\mu)$ and exponential stabilization intervals $S \sim \exp(\theta)$, the process $\{Z(t)\}$ is a continuous-time Markov chain with the state space $\{0, 1, \ldots, r\}$ and transition rate matrix $Q = (Q_{jj'})$:*

$$Q_{jj'} = \begin{cases} \theta, & j \neq r, j' = r, \\ j\mu, & 1 \le j \le r, j' = j - 1, \\ -\theta - j\mu, & j' = j < r, \\ -j\mu, & j = j' = r, \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

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

Proof. We first consider state $Z(t) = r$, i.e., the full list of successors at time t (see Fig. 3). Note that if a stabilization occurs when the current state is $Z(t) = r$, some current successors may be replaced by newly arriving users based on the successor rule. However, the successor failure rate is $\mu = 1/E[L]$ for both old successors and newly joining users due to the memoryless property of exponential distributions. Thus, it makes no difference whether new successors replace old ones or not (i.e., no matter if stabilizations happen when the state is r). This immediately follows that the transition probability from state r to $r - 1$ is $p_{r,r-1} = 1$, triggered by the failure of a successor, and the sojourn time in state r is exponential with rate $a_r = r\mu$. We then readily obtain that the transition rate from r to $r - 1$ is $a_r p_{r,r-1} = r\mu$.

Likewise, given that the stabilization intervals $S \sim \exp(\theta)$, it is not hard to obtain that the transition rate from state j to $j - 1$ is $j\mu$ for $1 \le j < r$, and the transition rate from state j to r is θ for $1 \le j < r$. This directly leads to the desired result. □

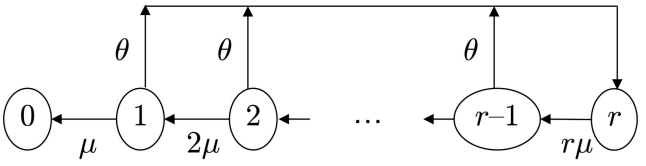

Fig. 3. Markov chain $\{Z(t)\}$ modeling a node's successor list.

The state diagram and transition rates of process $\{Z(t)\}$ are illustrated in Fig. 3, where each state models the number of alive successors and absorbing state 0 corresponds to user isolation. We usually write matrix Q in (21) in the canonical form:

$$Q = \begin{pmatrix} 0 & 0 \\ \mathbf{r} & Q_0 \end{pmatrix}, \tag{22}$$

where $\mathbf{r} = (q_{j0})^T$ for $j \neq 0$ is a column vector representing the transition rates to the absorbing state 0 and Q_0 is the rate matrix obtained by removing the rows and columns corresponding to state 0 from Q.

Define the first-hitting time T onto state 0 as

$$T = \inf(t > 0 : Z(t) = 0 | Z(0) = r\}. \tag{23}$$

Then, the isolation probability $\phi = P(T < L)$ can be reduced to [34, Theorem 11]:

$$\phi = \pi(0) V B V^{-1} \mathbf{r}, \tag{24}$$

where $\pi(0) = (0, \ldots, 1)_{1\times r}$ is the initial state distribution, V is a matrix of eigenvectors of Q_0, and $B = \mathrm{diag}(b_j)$ is a diagonal matrix with:

$$b_j = 1/(\mu - \xi_j), \tag{25}$$

$\mu = 1/E[L]$, $\xi_j \leq 0$ is the jth eigenvalue of Q_0, and Q_0 and $\mathbf{r}$ are in (22).

Simulation results of isolation probability ϕ are shown in Fig. 4. Notice from this figure that model (24) is very accurate compared to simulations. Also, observe that as ρ or r increase, node isolation probability sharply decreases. While (24) allows easy numerical computation, it provides little qualitative information about how ϕ behaves as a function of ρ and r. It is further difficult to compare the various stabilization strategies (studied later in the paper) if an explicit model of ϕ is not derived. We perform this task next.

4.3 Closed-Form Bounds on ϕ

Note from Fig. 2 that the sequence of stabilization intervals forms a renewal process with cycle length S. It then follows that isolation probability ϕ is equal to the probability that r successors simultaneously fail in *any* interval S before user v's lifetime expires. Note that the probability that all r successors fail in a particular interval S is given by

$$f = P(\max\{L_1, \ldots, L_r\} < S), \tag{26}$$

where $L_i \sim \exp(\mu)$ is the remaining lifetime of the ith successor at the beginning of a particular interval. Then, from Jensen's inequality [13, page 118], it is not hard to obtain the following closed-form upper bound on ϕ and prove that it becomes exact as the ratio $E[L]/E[S] \to \infty$.

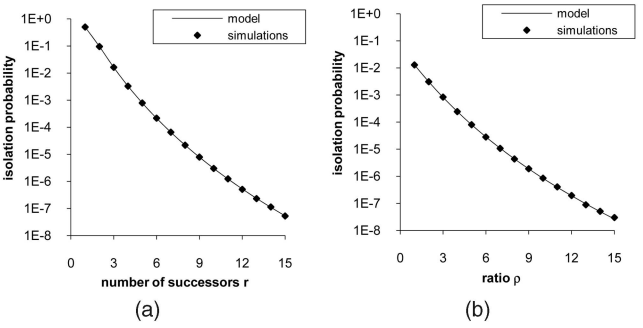

Fig. 4. Comparison of model (24) to simulation results on node isolation probability ϕ for exponential lifetimes with $E[L] = 0.5$ hours and exponential stabilization intervals with $E[S] = E[L]/\rho$. (a) fixed $\rho = 15$. (b) fixed $r = 16$.

Theorem 2. *For $L \sim \exp(\mu)$ and $S \sim \exp(\theta)$, isolation probability ϕ is upper-bounded by*

$$\phi < \rho f, \tag{27}$$

where $f = \rho! r!/(\rho + r)!$ and $\rho = E[L]/E[S] = \theta/\mu$. Moreover, the bound becomes tight as stabilization intervals become negligible compared to user lifetimes:

$$\lim_{\rho\to\infty} \frac{\phi}{\rho f} = 1. \tag{28}$$

Proof. Given that $S \sim \exp(\theta)$, probability f that all r successors fail with a particular interval S in (26) reduces to

$$f = \int_0^\infty \left(1 - e^{-\mu t}\right)^r \theta e^{-\theta t} dt. \tag{29}$$

Setting $\rho = \theta/\mu$ and $z = 1 - e^{-\mu t}$, (37) yields

$$f = \rho\mu \int_0^1 z^r (1-z)^\rho \frac{1}{\mu(1-z)} dz = \frac{\rho! r!}{(\rho + r)!}. \tag{30}$$

It is ready to see from (30) that as $\rho \to \infty$ and/or $r \to \infty$, $f \to 0$.

Next, note from Fig. 2 that the evolution of node v's successor list can be decomposed into a sequence of stabilization intervals. Let random variable D be the number of stabilization intervals with user v's lifetime L. Conditioning on $D = j$, we obtain that isolation probability $\phi(j)$ is approximately:

$$\phi(j) = 1 - (1 - f)^j, \quad j \geq 1, \tag{31}$$

where $(1 - f)^j$ is the probability that user v survives all j stabilization intervals and f is given in (30).

It is then clear from Jensen's inequality [13] in the discrete form that for concave function $\phi(j)$ shown in (31), the unconditional isolation probability ϕ yields:

$$\phi = E[\phi(D)] \leq 1 - (1 - f)^{E[D]}, \tag{32}$$

showing that our remaining task is to obtain $E[D]$.

For exponential S, it is not hard to obtain that the renewal function $E[D(t)]$, the expected number of stabilizations that have been executed by fixed time t, is simply:

$$E[D(t)] = \theta t, \quad \text{for all } t \geq 0. \tag{33}$$

Then, the mean number of stabilization intervals within random time units L can be obtained as

$$E[D] = \int_0^\infty E[D(t)] f_L(t) dt, \tag{34}$$

where $f_L(t)$ is the PDF of user lifetimes L. Substituting (33) into the above readily leads to

$$E[D] = \theta E[L] = \rho, \tag{35}$$

where $\rho = E[L]/E[S]$. Using (35), (32) can be transformed into

$$\phi \leq 1 - (1-f)^\rho \leq \rho f, \tag{36}$$

where $f < 1$ is given in (30), showing that ρf is an upper bound for ϕ.

Finally, note from Taylor expansion that as $\rho \to \infty$, $(1-f)^j \to 1 - jf$ for given j where $f = O(\rho^{-r})$ from (30). This immediately leads $\phi(j)$ in (31) into:

$$\frac{\phi(j)}{jf} = \frac{1-(1-f)^j}{jf} \to 1, \quad \rho \to \infty. \tag{37}$$

Invoking (37), isolation probability ϕ can be transformed into the following for $\rho \to \infty$:

$$\frac{\phi}{f\rho} = \frac{\sum_{j=1}^\infty \phi(j) P(D=j)}{f\rho} \to \frac{fE[D]}{f\rho}, \tag{38}$$

which directly leads to (28) recalling (35). □

The result in (28) indicates that for $\rho \to \infty$, probability ϕ for any user v to become isolated within its lifetime L can be approximated as the summation of probabilities that v is isolated in each stabilization interval. Indeed, an average user has approximately $\rho = E[L]/E[S]$ intervals in its lifetime and it gets isolated in any interval with probability f. Thus, since ϕ is asymptotically equal to ρf, isolation events in different intervals behave as if they were independent.

Table 2 illustrates the relative distance between the upper bound in (27) and the exact result (24) for $E[L] = 0.5$ hours and $r = 8$. It is clear from the table that as ρ increases, the two models converge and the upper bound is never violated. Also, note that other comparisons for different values of $E[L]$ and r exhibit similar results and are omitted for brevity.

We finish this section by examining how individual node isolations affect the connectivity of Chord as users continuously join and depart the system.

4.4 Graph Disconnection

Notice that Bernoulli variable X_i in (6) can be used to indicate whether user i is isolated due to the failure of its successor list under churn as well. Then, node isolation probability can be expressed as

TABLE 2
Comparison of the Asymptotic Model (27) to the Exact Model (24) of Node Isolation Probability ϕ with $E[L] = 0.5$ Hours, $\rho = E[L]/E[S]$, and $r = 8$

ρ	$E[S]$ s	exact model	upper bound	Relative Error
10	180	1.46×10^{-4}	2.29×10^{-4}	57.05%
50	36	2.30×10^{-8}	2.61×10^{-8}	13.41%
100	18	2.66×10^{-10}	2.84×10^{-10}	6.85%
200	9	2.55×10^{-12}	2.64×10^{-12}	3.46%
500	3.6	4.74×10^{-15}	4.80×10^{-15}	1.29%
1,000	1.8	3.86×10^{-17}	3.89×10^{-17}	0.69%

$$\phi = P(X_i = 1) = 1 - P(X_i = 0), \tag{39}$$

where ϕ is given by (24) or approximated by the upper bound in (27). If user i is isolated during its lifetime, we consider the system disconnected during that user's presence in the system; otherwise, the network is said to *survive* the join of peer i.

Supposing that N users have joined the system, we have that

$$X_N = \sum_{i=1}^N X_i \tag{40}$$

is the number of isolations among N join events. In the following, we use the Chen-Stein method [2] to study the probability that Chord survives N user joins without disconnection, i.e., $P(X_N = 0)$. Note that, again, this result is stronger than that in [17] since it applies to successor lists that exhibit much higher dependency during failure than neighbor lists studied in prior work and relies on more rigorous derivations.

Theorem 3. *Given that $N\phi r \to 0$ as $N \to \infty$, the probability that Chord survives N user joins without disconnection approaches:*

$$\lim_{N\to\infty} \frac{P(X_N = 0)}{(1-\phi)^N} = 1, \tag{41}$$

where X_N is defined in (40) and ϕ is given in (24).

Proof. The basic idea of the Chen-Stein method is that the distance between the distribution of X_N, i.e., a sum of N-dependent Bernoulli variables, and that of a Poisson random variable of the same mean can be upper-bounded by [2]

$$|P(X_N = 0) - P(V_N = 0)| \leq \alpha(b_1 + b_2 + b_3), \tag{42}$$

where V_N is a Poisson random variable with mean $E[V_N] = E[X_N] = N\phi$ and $\alpha = \min(1, 1/E[X_N])$ is no greater than 1. Convergence to the Poisson distribution happens when all of $b_1 - b_3$ tend to zero as $N \to \infty$. Our main task is to compute these metrics and observe under what condition they become negligibly small.

Define B_i to be a set of users who share at least one successor of user i in Chord:

$$B_i = \{i - r + 1, \ldots, i, \ldots, i + r - 1\}, \tag{43}$$

with $i \in B_i$ and size $|B_i| = 2r - 1$. Note that B_i represents the neighborhood of i such that variables X_i and X_j, for any node $j \in B_i$, are dependent. Metrics b_1 and b_2 are, respectively, given by [2]

$$b_1 := \sum_{i=1}^{N} \sum_{j \in B_i} P(X_i = 1)P(X_j = 1), \tag{44}$$

$$b_2 := \sum_{i=1}^{N} \sum_{j \neq i, j \in B_i} P(X_i = X_j = 1). \tag{45}$$

As in [2], $b_3 = 0$ since X_i is independent of X_j for all nodes j outside i's neighborhood B_i.

To calculate b_1, note that (44) yields

$$b_1 = \sum_{i=1}^{N} \sum_{j \in B_i} \phi^2 = N(2r-1)\phi^2. \tag{46}$$

Likewise, we obtain from (45) that

$$\begin{aligned} b_2 &= \sum_{i=1}^{N} \phi \sum_{j \neq i, j \in B_i} P(X_j = 1 | X_i = 1) \\ &\leq N\phi(2r-2). \end{aligned} \tag{47}$$

The last step is to observe that $b_1 = N\phi^2(2r-1) \to 0$ and $b_2 \leq N\phi(2r-2) \to 0$ as $N \to \infty$. Finally, given $b_1 + b_2 \to 0$, it is shown in (42) that X approaches a Poisson random variable with mean $E[X_N]$. This directly leads to

$$\lim_{N \to \infty} \frac{P(X_N = 0)}{e^{-E[X_N]}} = \lim_{N \to \infty} \frac{P(X_N = 0)}{e^{-N\phi}} = 1. \tag{48}$$

Recalling that $\phi \to 0$ as $N \to \infty$ given the assumption of this theorem and using Taylor expansion $e^{-\phi} = 1 - \phi$ for $\phi \to 0$, (47) yields

$$\lim_{N \to \infty} \frac{P(X_N = 0)}{(1-\phi)^N} = 1, \tag{49}$$

which establishes the desired result. □

Theorem 3 indicates that as long as ϕ is sufficiently small, probability $P(X_N = 0)$ that Chord accommodates N joining users without partitioning simply converges to the product of probabilities that individual nodes remain nonisolated. Observe that (41) holds under a wider set of conditions on ϕ that do not necessarily require $N\phi r \to 0$, but derivations in those cases are more tedious. Also, note that a typical way of accomplishing $N\phi r \to 0$ is to scale r with N so as to converge ϕ to zero faster than product Nr converges to infinity. The last note is that (41) is valid for any user lifetime distribution, where ϕ should be derived accordingly.

Armed with (41), we propose the following approximation to $P(X_N = 0)$ for finite N:

$$P(X_N = 0) \approx (1 - \phi)^N, \tag{50}$$

where the exact model of ϕ for exponential lifetimes is given by (24) and its asymptotic approximation is shown in (27).

Comparison of simulation results of $P(X_N = 0)$ to (50) is presented in Table 3 where model ϕ is computed based on (24). Notice from the first three columns in this table that simulation results are very close to (50) from $N = 10^3$ to 10^6 for $\rho = 40$. The rest of this table shows that as ρ increases (i.e., ϕ gets closer to zero), the model becomes more accurate as expected. Simulations for different r show similar results that are omitted for brevity. As an example of applying (50), assume that Chord has a mean size 5,000 users, $r = \lceil \log_2 5{,}000 \rceil = 13$ successors, $E[L] = 0.5$ hours, and $E[S] = 21$ seconds. We then obtain from (50) that the probability that Chord survives $N = 1$ billion user joins without disconnection is 0.999987. If we assume that each user joins and departs the network once per hour, this duration corresponds to 228 years. Furthermore, the system survives for $N = 100$ billion joins (i.e., 22,831 years) with probability 0.998558.

TABLE 3
Comparison of Model (50) of $P(X_N = 0)$ to Simulation Results for $r = 8$, Mean System Size 2,500, Exponential L with $E[L] = 0.5$ Hours, and Exponential S with $E[S] = E[L]/\rho$

$\rho = 40$ ($E[S] = 45$ s)			$N = 50{,}000$			
N	Simul.	(50)	ρ	$E[S]$ s	Simul.	(50)
1,000	1.000	.9999	16	112.5	.4831	.4557
5,000	.9996	.9995	24	75.0	.9176	.9139
8,000	.9993	.9993	32	56.3	.9833	.9829
10,000	.9992	.9991	40	45.0	.9954	.9955
50,000	.9954	.9955	48	37.5	.9985	.9985
100,000	.9910	.9910	56	32.1	.9995	.9994
500,000	.9555	.9556	64	28.1	.9998	.9998
1,000,000	.9129	.9131	80	25.7	1.000	.9999

5 Dynamic Node Failure: Effect of Stabilization Intervals

Results in the previous section only apply to exponential intervals S between two consecutive stabilizations. Though many modeling studies assume exponential stabilization intervals [14], [16] to obtain Markovian models, Chord by default uses uniform intervals [26]. In this section, we study isolation probability ϕ for uniform S, deal with ϕ for constant S, and then, find the optimal method for stabilizing successors.

5.1 Uniform Stabilization Delays

Denote by f_u the probability that all r successors of node v fail within interval S, where S is *uniformly* distributed in $[0, 2E[S]]$. Based on the renewal process with cycle length S, it is not hard to show that for uniform S, node isolation probability ϕ_u converges to

$$\frac{\phi_u}{\rho f_u} \to 1, \tag{51}$$

as $E[S] \to 0$, which is similar to the result shown in (28). Then, the ratio of isolation probability ϕ_u for uniform S to ϕ for exponential S is $\phi_u/\phi = f_u/f$, where f is given in (27). Deriving f_u, we obtain the next theorem.

Theorem 4. *For fixed r and $E[L]$ and uniform $S \in [0, 2E[S]]$, the ratio of isolation probability ϕ_u for uniform S to ϕ for exponential S converges to the following constant:*

$$\lim_{E[S]\to 0}\frac{\phi_u}{\phi}=\frac{2^r}{(r+1)!}. \quad (52)$$

Proof. The proof proceeds in two steps. First, for exponential L with a given $E[L]$ and uniform S in interval $[0, 2E[S]]$, f in (26) is reduced to

$$f_u=\int_0^\infty (1-e^{-\mu t})^r f_S(t)dt=\int_0^{2E[S]}\frac{(1-e^{-\mu t})^r}{2E[S]}dt.$$

Recalling $\rho=E[L]/E[S]$, the above yields

$$\begin{aligned} f_u&=\frac{\rho}{2}\int_0^{1-e^{-2/\rho}}\frac{x^r}{(1-x)}dx\\ &=\frac{\rho(1-e^{-2/\rho})^{r+1}}{2(r+1)}{}_2F_1(r+1,1;r+2;1-e^{-2/\rho}), \end{aligned}$$

where ${}_2F_1(a,b;c;z)$ is a hypergeometric function, which is always 1 for $z=0$. Note that as $E[S]\to 0$ (i.e., $\rho\to\infty$ since $E[L]$ is fixed), $z=1-e^{-2/\rho}\to 0$. This immediately follows that

$$\lim_{E[S]\to 0} f_u=\frac{\rho(1-e^{-2/\rho})^{r+1}}{2(r+1)}=\frac{2^r}{(r+1)\rho^r}, \quad (53)$$

where the last step is obtained using Taylor expansion.

Next, recall from (38) that isolation probability ϕ for any distribution of S can be expressed as the product of f and $E[D]$ as $\rho\to\infty$, where D is the random variable denoting the number of stabilization intervals with a lifetime L.

To obtain $E[D]$ for uniform S, we first derive $D(t)$ conditioning on user v's lifetime $L=t$. As $E[S]\to 0$ (which implies that $D(t)\to\infty$), it is clear from the strong law of large numbers that

$$D(t)E[S]\to t. \quad (54)$$

Invoking (53) and integrating $D(t)$ using PDF $f_L(t)$ of user lifetimes L leads to

$$E[S]E[D]=\int_0^\infty E[S]D(t)f_L(t)dt\to E[L], \quad (55)$$

as $E[S]\to 0$. The above can be easily transformed into

$$\lim_{E[S]\to 0}\frac{E[D]}{\rho}=1, \quad (56)$$

for any distribution of S. Combining (38) and (55), we immediately obtain isolation probability ϕ_u for uniform S:

$$\frac{\phi_u}{\rho f_u}\to 1,\quad E[S]\to 0. \quad (57)$$

It is then ready to see that the ratio of ϕ_u to ϕ shown in (28) for exponential S converges to

$$\frac{\phi_u}{\phi}\to\frac{f_u}{f},\quad \rho\to\infty, \quad (58)$$

where f is given in (27) and $\rho\to\infty$ is met under given assumptions in this theorem. Using Sterling's formula for $\rho\to\infty$ and fixed r, f in (27) can be reduced to

TABLE 4
Convergence of Simulation Results to Model $\phi_u/\phi=.0127$ from (52) for $E[L]=0.5$ Hours, $r=6$, and $\rho=E[L]/E[S]$

ρ	$E[S]$ s	Simulations of ϕ_u	Simulations of ϕ	ϕ_u/ϕ
20	90	2.15×10^{-6}	7.10×10^{-5}	.0303
40	45	7.59×10^{-8}	3.86×10^{-6}	.0197
60	30	9.98×10^{-9}	6.10×10^{-7}	.0164
80	22.5	2.28×10^{-9}	1.62×10^{-7}	.0141
100	18	7.18×10^{-10}	5.59×10^{-8}	.0128

$$\lim_{\rho\to\infty} f=r!\frac{e^r}{\rho^r}\left(1-\frac{r}{\rho+r}\right)^{\rho+r+1/2}=\frac{r!}{\rho^r}, \quad (59)$$

where the last step is obtained based on Taylor expansion for fixed r. Finally, substituting (53) and (59) into (58) directly leads to (52). □

Simulation results of ϕ_u for uniform S are shown in Table 4. Notice from this table that the ratio ϕ_u/ϕ indeed approaches that given by our model (52) as $E[S]$ becomes small. Since $\phi_u\le\phi$ for all r, the above result demonstrates that using uniform S is a better strategy than using exponential S and the amount of improvement becomes more significant when r increases, e.g., $\phi_u/\phi=7.055\times 10^{-4}$ for $r=8$ and $\phi_u/\phi=6.578\times 10^{-7}$ for $r=12$.

5.2 Constant Stabilization Delays

Next, following the derivations of ϕ_u/ϕ in Theorem 4, we easily obtain isolation probability ϕ_c for constant S.

Theorem 5. *For fixed r and $E[L]$ and constant S, the ratio of isolation probability ϕ_c to ϕ approaches:*

$$\lim_{E[S]\to 0}\frac{\phi_c}{\phi}=\frac{1}{r!}. \quad (60)$$

Proof. Following the derivations in the proof for Theorem 4, we readily obtain

$$f_c=(1-e^{-1/\rho})^r\to\rho^{-r},\quad \rho\to\infty, \quad (61)$$

and

$$\frac{\phi_c}{\phi}\to\frac{f_c}{f},\quad \rho\to\infty, \quad (62)$$

where f for exponential S is given in (59) and $\rho\to\infty$ is satisfied under given assumptions. Substituting (59) and (61) into (62) immediately leads to (60). □

Table 5 presents simulation results on ϕ_c when stabilization intervals are constant. Note that ratio ϕ_c/ϕ obtained from simulations is very close to that predicted by model (60) even for $\rho=60$ and that it converges to (60) as ρ increases further. Model (60) indicates that simply stabilizing successors at constant intervals can reduce isolation probability ϕ_c by a factor of $r!$ compared to ϕ as $E[S]\to 0$. To show the exact improvement over exponential S, we have $\phi_c/\phi=2.480\times 10^{-5}$ for $r=8$ and 2.088×10^{-9} for $r=12$. In addition, it is easy to notice from (51) and (59) that $\phi_c\le\phi_u$ and the ratio ϕ_c/ϕ_u approaches $(r+1)/2^r\le 1$ as $E[S]\to 0$. This ratio is 0.035 for $r=8$ and 0.003 for $r=12$.

TABLE 5
Convergence of Simulation Results to Model $\phi_c/\phi = .0014$ from (60) for $E[L] = 0.5$ Hours, $r = 6$, and $\rho = E[L]/E[S]$

ρ	$E[S]$ s	Simulations of ϕ_c	Simulations of ϕ	ϕ_c/ϕ
20	90	2.72×10^{-7}	7.10×10^{-5}	.0038
40	45	8.51×10^{-9}	3.86×10^{-6}	.0022
60	30	9.82×10^{-10}	6.10×10^{-7}	.0016
80	22.5	2.35×10^{-10}	1.62×10^{-7}	.0015
100	18	7.61×10^{-11}	5.59×10^{-8}	.0014

5.3 Optimal Strategy

The above analysis shows that for exponential lifetimes, the ratio of ϕ_c under constant S to ϕ_o under any other S can be transformed into

$$\lim_{E[S]\to 0} \frac{\phi_c}{\phi_o} = \frac{P(\max\{L_1, \ldots, L_r\} < E[S])}{P(\max\{L_1, \ldots, L_r\} < S)}, \tag{63}$$

where $L_i \sim \exp(\mu)$ is the residual lifetime of the ith successor of node v at the beginning of a particular interval. While we already established that the above ratio is asymptotically less than 1 for both exponential and uniform S, the next theorem indicates that the same result holds for all other distributions as well.

Theorem 6. *For exponential user lifetimes with fixed $E[L] > 0$ and the same mean stabilization interval $E[S] \to 0$, node isolation probability ϕ_c under constant S is no greater than that under any random S.*

Proof. For exponential user lifetimes with mean $E[L] = 1/\mu$, recall that the probability that all r successors of node v fail within a particular interval S is

$$P(\max\{L_1, \ldots, L_r\} < S) = \int_0^\infty G(x) f_S(x) ds, \tag{64}$$

where $G(x) = P(\max\{L_1, \ldots, L_r\} < x) = (1 - e^{-\mu x})^r$. The second derivative of $G(x)$ is thus

$$G''(x) = r\mu^2 e^{-\mu x}(1 - e^{-\mu x})^{r-2}(re^{-\mu x} - 1), \tag{65}$$

for $r \geq 3$. Then, it is easy to see that for $r \geq 3$:

$$\begin{cases} G''(x) > 0, & x < E[L] \ln r, \\ G''(x) \leq 0, & \text{otherwise}, \end{cases} \tag{66}$$

which indicates that $G(x)$ is a convex function for $x < E[L] \ln r$ and concave for $x > E[L] \ln r$.

For $E[S] \to 0$, note that $S \leq E[L] \ln r$ holds with probability approaching 1. This immediately transforms (64) into

$$P(\max\{L_1, \ldots, L_r\} < S) = \int_0^{E[L]\ln r} G(x) f_S(x) ds, \tag{67}$$

showing that the convex part of $G(x)$ determines the above metric. Then, for $E[S] \to 0$, we obtain from Jensen's inequality [13] that

$$P(\max\{L_1, \ldots, L_r\} < S) \geq P(\max\{L_1, \ldots, L_r\} < E[S]),$$

since $G(x)$ is strictly convex for $x < E[L] \ln r$. This directly leads to

$$\lim_{E[S]\to 0} \frac{\phi_c}{\phi_o} = \frac{P(\max\{L_1, \ldots, L_r\} < E[S])}{P(\max\{L_1, \ldots, L_r\} < S)} \leq 1, \tag{68}$$

for any random S, which completes the proof. □

Theorem 6 shows that using constant S is not only a simple, but also *optimal* method to stabilize successors in Chord.

5.4 Discussion

In this section, we have shown that constant stabilization intervals outperform any other distribution of S in terms of resilience against disconnection. This has a simple intuitive explanation. Recalling the successor list model in Fig. 2, observe that node isolation more likely occurs during the longest stabilization interval within a user lifetime. Therefore, by reducing the variance of S (i.e., keeping all intervals equally small), one achieves the lowest probability of disconnection, which is the rationale behind Theorem 6.

One straightforward approach that takes advantage of this finding is for each node v to perform the jth stabilization at time $t_v + j\Delta_v$, where t_v is the instance when v joins the system and Δ_v is its interstabilization delay, which may depend on the user's resources (CPU, memory), bandwidth, and desire to stay connected. In situations where users may simultaneously arrive with similar Δ_v, a random initial shift may be used to prevent synchronization effects. In other words, the jth stabilization may occur at time $t_v + \omega_v + j\Delta_v$, where ω_v is some small initial delay (generated by each user upon join).

6 Heavy-Tailed Lifetimes

Without the memoryless property on lifetime L, derivation of probability f that all r successors fail within interval S is simply intractable. However, for systems with heavy-tailed lifetimes [5], [30] where old users are more likely to remain alive for a longer time in the system, a mixture of old and new users within a given successor list leads to a smaller f compared to that for exponential lifetimes. Thus, the probability of node isolation due to failure of the entire successor list in Chord is *smaller* when the distribution of user lifetimes is heavy-tailed compared to the exponential case studied earlier in this paper, which we next confirm in simulations.

We examine four different distributions of interval S, including exponential with rate $1/E[S]$, Pareto with CDF $F(x) = 1 - (1 + x/\beta)^{-\alpha}$ where $\alpha = 3$ and $\beta = (\alpha - 1)E[S]$, uniform in $[0, 2E[S]]$, and constant equal to $E[S]$. Simulation results of isolation probability ϕ for exponential and Pareto lifetimes under the four stabilization strategies are plotted in Fig. 5. Notice in the figure that S with the highest variance (i.e., Pareto S) performs the worst, followed by exponential and uniform cases, while constant S is the best. Further, observe that ϕ for Pareto lifetimes is smaller than that for exponential lifetimes under all four stabilization strategies and the difference becomes smaller as $E[S]$ decreases. In fact, the model is a very close match to the Pareto case in Figs. 5c and 5d. These observations confirm that our exponential model of ϕ provides an upper bound for systems with heavy-tailed lifetimes over a wide range of stabilization delays S.

Since graphs in Figs. 5a, 5b, 5c, and 5d are listed in the order of decaying variance $Var[S]$, the figure can be used to

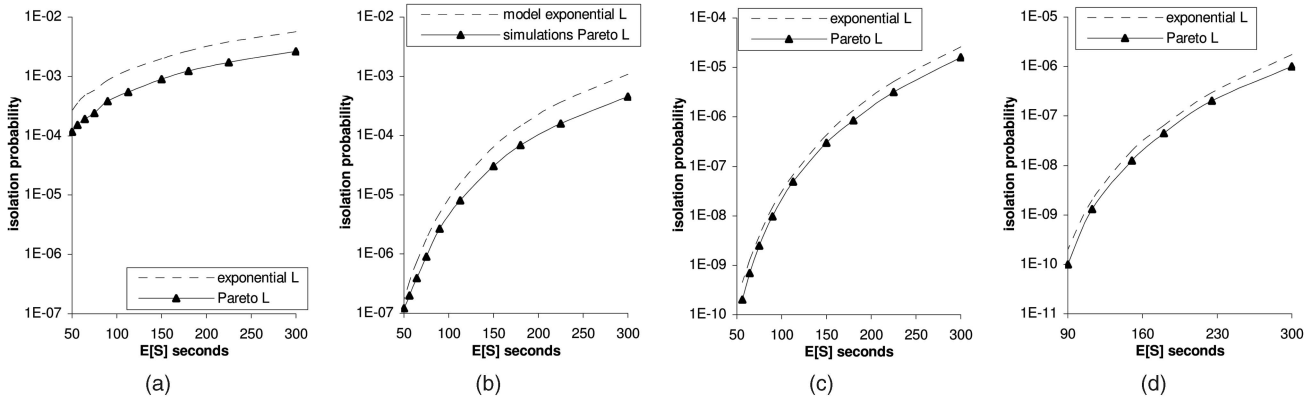

Fig. 5. Comparison of simulation results on node isolation probability ϕ under different stabilization strategies for exponential and Pareto lifetimes with $\alpha = 3$ and $E[L] = 0.5$ hours, mean system size 2,500, and $r = 8$ in Chord. (a) Pareto S. (b) Exponential S. (c) Uniform S. (d) Constant S.

suggest that larger variance $Var[S]$ leads to larger isolation probability ϕ not only for exponential lifetimes L, but also for Pareto. The next lemma formally establishes the result for Pareto lifetimes.

Lemma 2. *Let S_1 and S_2, respectively, denote random intervals of any two stabilization methods with the same mean $E[S_1] = E[S_2]$. For Pareto lifetimes with CDF $1-(1+x/\beta)^{-\alpha}$, $\alpha > 1$ and $\beta > 0$, the following holds:*

$$Var[S_1] \geq Var[S_2] \Longrightarrow \phi(S_1) \geq \phi(S_2), \tag{69}$$

where $\phi(S)$ represents isolation probability for stabilization intervals with random variable S.

Proof. Note from (26) that the probability that all of user i's r successors fail in the jth stabilization interval is

$$f(j) = \int_0^\infty P(\max(L_1^*, \ldots, L_r^*) \leq x) dF_S(x), j = 1, 2, \ldots,$$

where $F_S(x)$ is the CDF of stabilization intervals and L_k^* is the remaining lifetime of the kth successor from when the jth stabilization starts until this successor departs, $k = 1, \ldots, r$. Due to user independence, the above is reduced to

$$f(j) = \int_0^\infty \prod_{k=1}^{r} P(L_k^* < x) dF_S(x). \tag{70}$$

Defining $h(x) := \prod_{k=1}^{r} P(L_k^* < x)$, (70) yields

$$f(j) = \int_0^\infty h(x) dF_S(x) = E[h(S)]. \tag{71}$$

Next, observe that for Pareto lifetimes, the probability that a node's remaining lifetime L_k^* is less than x conditioned on that its age A is y is given by

$$P(L_k^* < x | A = y) = 1 - \left(1 + \frac{x}{\beta + y}\right)^{-\alpha}, \tag{72}$$

for $x \geq 0, y \geq 0$. In case of age $y = 0$, (72) is simply the lifetime CDF of a newly arrived user. Since the second derivative of (72) with respect to x is nonnegative for *any* given age $y \geq 0$, we reach that the second derivative of $h(x)$ is nonnegative, showing that $h(x)$ is a nondecreasing convex function.

We thus obtain from [32, p. 488] that for any nondecreasing convex function $h(x)$,

$$Var[S_1] \geq Var[S_2] \Longrightarrow E[h(S_1)] \geq E[h(S_2)],$$

which then leads to the desired result. □

Lemma 2 shows that for Pareto lifetimes, stabilization with constant intervals S (whose $Var[S] = 0$) leads to the smallest isolation probability, which is consistent with Theorem 6. However, this result may not hold for other distributions of user lifetimes since it depends on whether function $h(x)$ in (71) is convex or not.

7 Conclusion

This paper tackled the problem of deriving formulas for the resilience of Chord's successor list under both static and dynamic node failure. We found that under static node failure, Chord exhibited the same resilience through the successor list as that many other DHTs and unstructured P2P networks [17] through their randomized neighbor tables. We also showed that when Chord experienced continuous node joins/departures, node isolation could be treated as independent when system size and successor lists were sufficiently large. We further demonstrated that stabilization with constant intervals was optimal and kept Chord connected with the highest probability under exponential and Pareto lifetimes. Similar results are expected to hold for other heavy-tailed distributions and a variety of nonchord DHTs. We leave this direction for future work.

Acknowledgments

An earlier version of this paper appeared in IEEE INFOCOM 2008. This work is supported by US National Science Foundation (NSF) grant CNS-0720571.

References

[1] R. Albert and A. Barabási, "Topology of Evolving Networks: Local Events and Universality," *Physical Rev. Letters*, vol. 85, no. 24, pp. 5234-5237, Dec. 2000.

[2] R. Arratia, L. Goldstein, and L. Gordon, "Two Moments Suffice for Poisson Approximations: The Chen-Stein Method," *The Annals of Probability*, vol. 17, no. 1, pp. 9-25, Jan. 1989.

[3] J. Aspnes, Z. Diamadi, and G. Shah, "Fault-Tolerant Routing in Peer-to-Peer Systems," *Proc. ACM Symp. Principles of Distributed Computing (PODC),* pp. 223-232, July 2002.
[4] N. Balakrishnan and M.V. Koutras, *Runs and Scans with Applications.* John Wiley & Sons, 2002.
[5] F.E. Bustamante and Y. Qiao, "Friendships That Last: Peer Lifespan and Its Role in P2P Protocols," *Proc. Int'l Workshop Web Content Caching and Distribution,* Sept. 2003.
[6] M. Castro, M. Costa, and A. Rowstron, "Performance and Dependability of Structured Peer-to-Peer Overlays," *Proc. Int'l Conf. Dependable Systems and Networks (DSN),* June 2004.
[7] B.-G. Chun, B. Zhao, and J. Kubiatowicz, "Impact of Neighbor Selection on Performance and Resilience of Structured P2P Networks," *Proc. Int'l Workshop Peer-to-Peer Systems (IPTPS),* pp. 264-274, Feb. 2005.
[8] H. Frank, "Maximally Reliable Node Weighted Graphs," *Proc. Third Ann. Conf. Information Sciences and Systems,* pp. 1-6, Mar. 1969.
[9] P.B. Godfrey, S. Shenker, and I. Stoica, "Minimizing Churn in Distributed Systems," *Proc. ACM SIGCOMM,* Sept. 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," *Proc. ACM SIGCOMM,* pp. 381-394, Aug. 2003.
[11] M.F. Kaashoek and D. Karger, "Koorde: A Simple Degree-Optimal Distributed Hash Table," *Proc. Int'l Workshop Peer-to-Peer Systems (IPTPS),* pp. 98-107, Feb. 2003.
[12] A.K. Kelmans, "Connectivity of Probabilistic Networks," *Automation and Remote Control,* vol. 29, pp. 444-460, 1967.
[13] S.G. Krantz, *Handbook of Complex Variables.* Birkhäuser, 1999.
[14] S. Krishnamurthy, S. El-Ansary, E. Aurell, and S. Haridi, "A Statistical Theory of Chord under Churn," *Proc. Int'l Workshop Peer-to-Peer Systems (IPTPS),* pp. 93-103, Feb. 2005.
[15] S.S. Lam and H. Liu, "Failure Recovery for Structured P2P Networks: Protocol Design and Performance Evaluation," *Proc. ACM SIGMETRICS,* pp. 199-210, June 2004.
[16] D. Leonard, V. Rai, and D. Loguinov, "On Lifetime-Based Node Failure and Stochastic Resilience of Decentralized Peer-to-Peer Networks," *Proc. ACM SIGMETRICS,* pp. 26-37, June 2005.
[17] D. Leonard, Z. Yao, X. Wang, and D. Loguinov, "On Static and Dynamic Partitioning Behavior of Large-Scale Networks," *Proc. IEEE Int'l Conf. Network Protocols (ICNP),* pp. 345-357, Nov. 2005.
[18] J. Li, J. Stribling, T.M. Gil, R. Morris, and M.F. Kaashoek, "Comparing the Performance of Distributed Hash Tables under Churn," *Proc. Int'l Workshop Peer-to-Peer Systems (IPTPS),* pp. 87-99, Feb. 2004.
[19] 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," *Proc. IEEE INFOCOM,* pp. 225-236, Mar. 2005.
[20] D. Liben-Nowell, H. Balakrishnan, and D. Karger, "Analysis of the Evolution of the Peer-to-Peer Systems," *Proc. ACM Symp. Principles of Distributed Computing (PODC),* pp. 233-242, July 2002.
[21] P. Maymounkov and D. Mazieres, "Kademlia: A Peer-to-Peer Information System Based on the XOR Metric," *Proc. Int'l Workshop Peer-to-Peer Systems (IPTPS),* pp. 53-65, Mar. 2002.
[22] G. Pandurangan, P. Raghavan, and E. Upfal, "Building Low-Diameter Peer-to-Peer Networks," *IEEE J. Selected Areas in Comm.,* vol. 21, no. 6, pp. 995-1002, Aug. 2003.
[23] L. Plissonneau, J.-L. Costeux, and P. Brown, "Analysis of Peer-to-Peer Traffic on ADSL," *Proc. Passive and Active Measurement Conf. (PAM),* pp. 69-82, Mar. 2005.
[24] S. Rhea, D. Geels, T. Roscoe, and J. Kubiatowicz, "Handling Churn in a DHT," *Proc. USENIX Ann. Technical Conf.,* pp. 127-140. June 2004.
[25] A. Rowstron and P. Druschel, "Pastry: Scalable, Decentralized Object Location and Routing for Large-Scale Peer-to-Peer Systems," *Proc. IFIP/ACM Int'l Conf. Distributed Systems Platforms (Middleware),* pp. 329-350, Nov. 2001.
[26] I. Stoica, R. Morris, D. Liben-Nowell, D.R. Karger, M.F. Kaashoek, F. Dabek, and H. Balakrishnan, "Chord: A Scalable Peer-to-Peer Lookup Protocol for Internet Applications," *IEEE/ACM Trans. Networking,* vol. 11, no. 1, pp. 17-32, Feb. 2003.
[27] D. Stutzbach and R. Rejaie, "Understanding Churn in Peer-to-Peer Networks," *Proc. ACM Internet Measurement Conf. (IMC),* pp. 189-202, Oct. 2006.
[28] K. Sutner, A. Satyanarayana, and C. Suffel, "The Complexity of the Residual Node Connectedness Reliability Problem," *SIAM J. Computing,* vol. 20, pp. 149-155, 1991.
[29] G. Tan and S. Jarvis, "Stochastic Analysis and Improvement of the Reliability of DHT-Based Multicast," *Proc. IEEE INFOCOM,* pp. 2198-2206, May 2007.
[30] X. Wang, Z. Yao, and D. Loguinov, "Residual-Based Estimation of Peer and Link Lifetimes in P2P Networks," *IEEE/ACM Trans. Networking,* vol. 17, no. 3, pp. 726-739, June 2009.
[31] X. Wang, Z. Yao, and D. Loguinov, "Residual-Based Measurement of Peer and Link Lifetimes in Gnutella Networks," *Proc. IEEE INFOCOM,* pp. 391-399, May 2007.
[32] R.W. Wolff, *Stochastic Modeling and the Theory of Queues.* Prentice Hall, 1989.
[33] Z. Yao, X. Wang, D. Leonard, and D. Loguinov, "Node Isolation Model and Age-Based Neighbor Selection in Unstructured P2P Networks," *IEEE/ACM Trans. Networking,* vol. 17, no. 2, pp. 144-157, Apr. 2009.
[34] Z. Yao, D. Leonard, X. Wang, and D. Loguinov, "Modeling Heterogeneous User Churn and Local Resilience of Unstructured P2P Networks," *Proc. IEEE Int'l Conf. Network Protocols (ICNP),* pp. 32-41, Nov. 2006.

Zhongmei Yao received the BS degree in engineering from Donghua University, Shanghai, China, in 1997, the MS degree in computer science from Louisiana Tech University, Ruston, in 2004, and the PhD degree in computer science from Texas A&M University, College Station, in 2009. Since 2009, she has been an assistant professor in the Department of Computer Science at The University of Dayton, Ohio. Her research interests include peer-to-peer systems, Internet measurement, and stochastic network modeling. She is a member of the IEEE.

Dmitri Loguinov (S '99-M '03-SM '08) received the BS degree (with honors) in computer science from Moscow State University, Russia, in 1995, and the PhD degree in computer science from the City University of New York, in 2002. Between 2002 and 2007, he was an assistant professor in the Department of Computer Science at Texas A&M University, College Station, where he is currently an associate professor and the director of the Internet Research Lab (IRL). His research interests include peer-to-peer networks, video streaming, congestion control, Internet measurement, and modeling. He is a senior member of the IEEE.