On Lifetime-Based Node Failure and Stochastic Resilience of Decentralized Peer-to-Peer Networks

Derek Leonard, Vivek Rai, and Dmitri Loguinov*
Computer Science Department
Texas A&M University, College Station, TX 77843 USA
{dleonard, vivekr, dmitri}@cs.tamu.edu

ABSTRACT

To understand how high rates of churn and random departure decisions of end-users affect connectivity of P2P networks, this paper investigates resilience of random graphs to lifetime-based node failure and derives the expected delay before a user is forcefully isolated from the graph and the probability that this occurs within his/her lifetime. Our results indicate that systems with heavy-tailed lifetime distributions are more resilient than those with light-tailed (e.g., exponential) distributions and that for a given average degree, k-regular graphs exhibit the highest resilience. As a practical illustration of our results, each user in a system with $n = 100$ billion peers, 30-minute average lifetime, and 1-minute node-replacement delay can stay connected to the graph with probability $1 - 1/n$ using only 9 neighbors. This is in contrast to 37 neighbors required under previous modeling efforts. We finish the paper by showing that many P2P networks are *almost surely* (i.e., with probability $1 - o(1)$) connected if they have no isolated nodes and derive a simple model for the probability that a P2P system partitions under churn.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Modeling techniques

General Terms

Algorithms, Performance, Theory

Keywords

Peer-to-Peer, Pareto, Stochastic Lifetime Resilience

1. INTRODUCTION

Resilience of random graphs [6] and various types of deterministic networks [7], [17] has attracted significant attention in research literature. A classical problem in this line of study is to understand failure conditions under which the network disconnects and/or starts to offer noticeably lower performance (such as increased routing distance) to its users. To this end, many existing models assume uniformly random edge/node failure and derive the conditions under which each user [33], certain components [6], or the entire graph [14] stay connected after the failure.

Current analysis of P2P networks frequently adopts the same model of uniform, concurrent node failure and ranges from single-node isolation [16], [33] to disconnection of the entire graph [4], [12], [14], [25]; however, these studies rarely discuss the scenarios under which large quantities of P2P users may simultaneously fail or how to accurately estimate failure probability p in practical systems such as KaZaA or Gnutella.[1] Another popular model assumes the existence of a rogue entity that compromises arbitrary nodes in the system [11], [30] and additional P2P studies examine such metrics as the required rate of neighbor replacement to avoid disconnection [22] and the delay before the system recovers from inconsistencies [26].

It has been recently noted [5] that realistic P2P failure models should take into account the intrinsic behavior of Internet users, who depart from the network based on a combination of factors often more complex than the traditional binary metric. In these networks, failure develops when users voluntarily decide to leave the system based on their attention span and/or browsing habits. To examine the behavior of such systems, this paper introduces a simple node-failure model based on user lifetimes and studies the resilience of P2P networks in which nodes stay online for random periods of time. In this model, each arriving user is assigned a random lifetime L_i drawn from some distribution $F(x)$, which reflects the behavior of the user and represents the duration of his/her services (e.g., forwarding queries, sharing files) to the P2P community.

We start our investigation with the *passive* lifetime model in which failed neighbors are not continuously replaced. It is interesting to observe that even in this case, a large fraction of nodes are capable of staying online for their entire lifespan without suffering an isolation. Furthermore, we show that depending on the tail-weight of the lifetime distribution, the probability of individual node isolation can be made arbitrarily small *without* increasing node degree. While the passive model certainly allows P2P networks to evolve as long as newly arriving nodes replenish enough broken links in the system, most practical P2P networks employ special neighbor-recovery strategies and attempt to repair the

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

SIGMETRICS'05, June 6–10, 2005, Banff, Alberta, Canada.

[1] In the absence of a better estimate, value $p = 1/2$ is often used for illustration purposes [33].

failed segments of the graph. We thus subsequently study the *active* model where each failed neighbor is replaced with another node after some random search delay. For this scenario, we derive both the expected time to isolation $E[T]$ and the probability π that this event occurs within the lifetime of a user.

In contrast to the p–percent failure model, the lifetime framework does not require estimation of such intricate metrics as the exact value of p or even the shape of the lifetime distribution. Towards the end of the paper, we show a reasonably good upper bound on π that only requires the mean user lifetime and the average node-replacement delay, both of which are easily measurable in existing systems.

Note that throughout most of the paper, "resilience" generally refers to the ability of an arriving user i to stay connected to the rest of the graph for duration L_i while its neighbors are constantly changing. This is arguably the most transparent and relevant metric from the end-user's perspective; however, to complete the picture we also address *global* resilience where we informally show that tightly connected graphs (such as DHTs and many k-regular random graphs) partition *with at least one isolated node* with probability $1 - o(1)$ as the size of the network $n \to \infty$. This result demonstrates that metric π solely determines the probability that an evolving P2P network partitions under churn and that global disconnection of such graphs for sufficiently small π almost surely involves a *single* node.

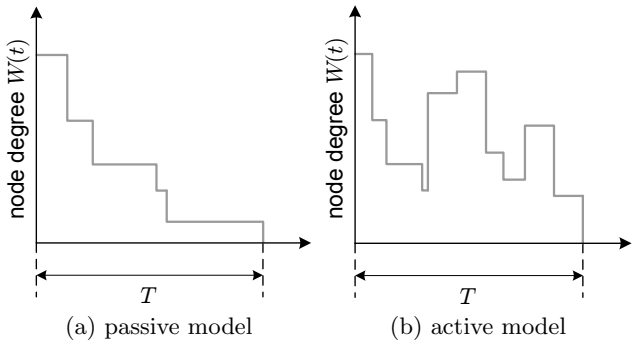
2. LIFETIME-BASED NODE FAILURE

In this section, we introduce our model of node failure and explain the assumptions used later in the paper.

2.1 Lifetime Model and Paper Overview

In the discussion that follows, we consider k-regular P2P graphs and analyze the probability that a randomly selected node v is forced to disconnect from the system because all of its neighbors have simultaneously departed and left it with no way to route within the graph. For each user i in the system, let L_i be the amount of time that the user stays in the network searching for content, browsing for information, or providing services to other peers. It has been observed that the distribution of user lifetimes in real P2P systems is often heavy-tailed (i.e., Pareto) [8], [31], where most users spend minutes per day browsing the network while a handful of other peers exhibit server-like behavior and keep their computers logged in for weeks at a time. To allow arbitrarily small lifetimes, we use a shifted Pareto distribution $F(x) = 1 - (1 + x/\beta)^{-\alpha}, x > 0, \alpha > 1$ to represent heavy-tailed user lifetimes, where scale parameter $\beta > 0$ can change the mean of the distribution without affecting its range $(0, \infty]$. Note that the mean of this distribution $E[L_i] = \beta/(\alpha - 1)$ is finite only if $\alpha > 1$, which we assume holds in the rest of the paper. While our primary goal is the study of human-based P2P systems, we also aim to keep our results universal and applicable to other systems of non-human devices and software agents where the nodes may exhibit non-Pareto distributions of L_i. Thus, throughout the paper, we allow a variety of additional user lifetimes ranging from heavy-tailed to exponential.

The most basic question a joining user may ask about the resilience of lifetime-based P2P systems is *what is the probability that I can outlive all of my original neighbors*? We call this model "passive" since it does not involve any neighbor replacement and study it in fair detail in the next section. This model arises when the search time S to find neighbor replacement is prohibitively high (i.e., significantly above $E[L_i]$) or when peers intentionally do not attempt to repair broken links. If degree k is sufficiently large, it is intuitively clear that a given node v is not likely to out-survive k other peers; however, it is interesting to observe that Pareto distributions of L_i make this probability significantly smaller compared to the "baseline" exponential case.



(a) passive model (b) active model

Figure 1: Degree evolution process leading to isolation under (a) passive and (b) active models.

In a later part of the paper, we allow users to randomly (with respect to the lifetime of other peers) search the system for new neighbors once the failure of an existing neighbor is detected. We call this model "active" to contrast the actions of each user with those in the passive model. Defining $W(t)$ to be the degree of v at time t, the difference between the passive and active models is demonstrated in Figure 1, which shows the evolution of $W(t)$ and the isolation time T for both models.

2.2 Modeling Assumptions

To keep the derivations tractable, we impose the following restrictions on the system. We first assume that v joins a network that has evolved sufficiently long so as to overcome any transient effects and allow asymptotic results from renewal process theory to hold. This assumption is usually satisfied in practice since P2P systems continuously evolve for hundreds of days or weeks before being restarted (if ever) and the average lifetime $E[L_i]$ is negligible compared to the age of the whole system when any given node joins it.

Our second modeling assumption requires certain stationarity of lifetime L_i. This means that users joining the system at different times of the day or month have their lifetimes drawn from the same distribution $F(x)$. While it may be argued that users joining late at night browse the network longer (or shorter) than those joining in the morning, our results below can be easily extended to non-stationary environments and used to derive upper/lower bounds on the performance of such systems.

Finally, we should note that these stationarity assumptions do not apply to the number of nodes n, which we allow to vary with time according to *any* arrival/departure process as long as $n \gg 1$ stays sufficiently large. We also allow arbitrary routing changes in the graph over time and are not concerned with the routing topology or algorithms used to forward queries. Thus, our analysis is applicable to both structured (i.e., DHTs) and unstructured P2P systems.

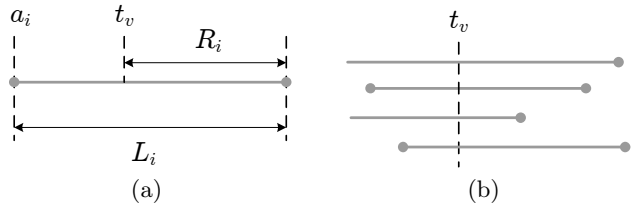


Figure 2: (a) A neighbor's lifetime $L_i = d_i - a_i$ and its residual life $R_i = d_i - t_v$. (b) Location of t_v is uniformly random with respect to the lives of all neighbors.

3. PASSIVE LIFETIME MODEL

We start by studying the resilience of dynamic P2P systems under the assumption that users do *not* attempt to replace the failed neighbors. As we show below, this analysis can be reduced to basic renewal process theory; however, its application to P2P networks is novel. Due to limited space, we omit the proofs of certain straightforward lemmas and refer the reader to the technical report [20].

3.1 Model Basics

We first examine the probability that a node v can outlive k randomly selected nodes if all of them joined the system *at the same time*. While the answer to this question is trivial, it provides a lower-bound performance of the system and helps us explain the more advanced results that follow.

LEMMA 1. *The probability that node v has a larger lifetime than k randomly selected nodes is $1/(k+1)$.*

Consider an example of Chord [33] with neighbor table size equal to $\log_2 n$, where n is the total number of nodes in the P2P network. Thus, in a system with 1 million nodes, the probability that a randomly selected node outlives $\log_2 n$ other peers is approximately 4.8%. This implies that with probability 95.2%, a user v *does not have to replace any of its neighbors to remain online for the desired duration L_v*.

Note, however, that in current P2P networks, it is neither desirable nor possible for a new node v to pick its neighbors such that their arrival times are exactly the same as v's. Thus, when v joins a P2P system, it typically must randomly select its k neighbors from the nodes *already present* in the network. These nodes have each been alive for some random amount of time before v's arrival, which may or may not affect the remainder of their online presence. In fact, the tail-weight of the distribution of L_i will determine whether v's neighbors are likely to exhibit longer or shorter remaining lives than v itself.

Throughout the paper, we assume that neighbor selection during join and replacement is *independent* of 1) neighbors' lifetimes L_i or 2) their current ages A_i. The first assumption clearly holds in most systems since the nodes themselves do not know how long the user plans to browse the network. Thus, the value of L_i is generally hard to correlate with any other metric (even under adversarial selection). The second assumption holds in most current DHTs [16], [27], [29], [33] and unstructured graphs [9], [13], [32] since neighbor selection depends on a variety of factors (such as a uniform hashing function of the DHT space [33], random walks [13], interest similarity [32], etc.), none of which are correlated with node age.

The above assumptions allow one to model the time when v selects each of its k neighbors to be uniformly random within each neighbor's interval of online presence. This is illustrated in Figure 2(a), where t_v is the join time of node v, and a_i and d_i are the arrival and departure times of neighbor i, respectively. Since the system has evolved for sufficiently long before v joined, the probability that v finds neighbor i at any point within the interval $[a_i, d_i]$ can be modeled as equally likely. This is schematically shown in Figure 2(b) for four neighbors of v, whose intervals $[a_i, d_i]$ are independent of each other or the value of t_v.

Next, we formalize the notion of residual lifetimes and examine under what conditions the neighbors are more likely to outlive each joining node v. Define $R_i = d_i - t_v$ to be the remaining lifetime of neighbor i when v joined the system. As before, let $F(x)$ be the CDF of lifetime L_i. Assuming that n is large and the system has reached stationarity, the CDF of residual lifetimes is given by [28]:

$$F_R(x) = P(R_i < x) = \frac{1}{E[L_i]} \int_0^x (1 - F(z))dz. \qquad (1)$$

For exponential lifetimes, the residuals are trivially exponential using the memoryless property of $F(x)$: $F_R(x) = 1 - e^{-\lambda x}$; however, the next result shows that the residuals of Pareto distributions with shape α are *more* heavy-tailed and exhibit shape parameter $\alpha - 1$.

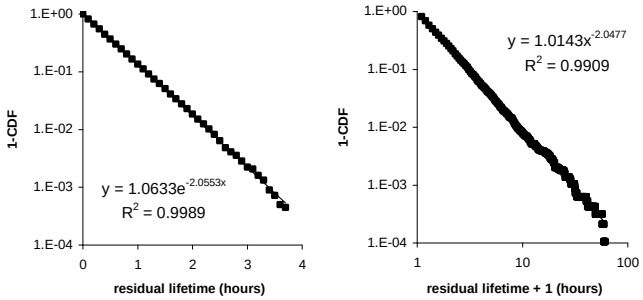
LEMMA 2. *The CDF of residuals for Pareto lifetimes with $F(x) = 1 - (1 + x/\beta)^{-\alpha}, \alpha > 1$ is given by:*

$$F_R(x) = 1 - \left(1 + \frac{x}{\beta}\right)^{1-\alpha}. \qquad (2)$$

This outcome is not surprising as it is well-known that heavy-tailed distributions exhibit "memory," which means that users who survived in the system for some time $t > 0$ are likely to remain online for longer periods of time than the arriving users. In fact, the larger the current age of a peer, the longer he/she is expected to remain online. The occurrence of this "heavy-tailed" phenomenon in P2P systems is supported by experimental observations [8] and can also be explained on the intuitive level. If a user v has already spent 10 hours in the system, it is generally unlikely that he/she will leave the network in the next 5 minutes; however, the same probability for newly arriving peers is substantially higher as some of them depart almost immediately [31].

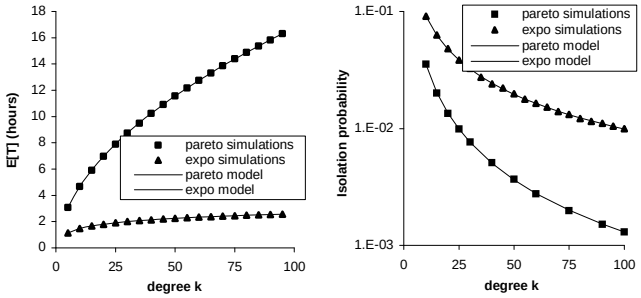
Since the rest of the derivations in the paper rely on (1), it is important to verify that asymptotic approximations from renewal process theory actually hold in practice. We created a hypothetical system with $n = 1000$ users and degree $k = 10$, in which each node lived for a random duration L_i and then departed from the system. To prevent the network size from depleting to zero, each failed node was immediately replaced by a fresh node with another random lifetime L_j (the exact arrival process was not essential and had no effect on the results). For each new arrival v into the system, we recorded the residual lifetimes of the neighbors that v randomly selected from the pool of $n - 1$ online peers.

Results of two typical simulations are plotted in Figure 3 for the exponential and Pareto lifetimes. As we often do throughout the paper, parameters α and λ are selected so that $E[L_i]$ is 0.5 hours for both distributions and the scaling parameter β is set to 1 in the Pareto $F(x)$. As the figure



(a) Exponential with $\lambda = 2$ (b) Pareto with $\alpha = 3, \beta = 1$

Figure 3: Residual lifetimes in simulation.



(a) $E[T]$ (b) π

Figure 4: Comparison of models (4)-(5) and (7)-(8) with simulation results .

shows, the residual exponential distribution remains exponential, while the Pareto case becomes more heavy-tailed and indeed exhibits shape parameter $\alpha - 1 = 2$. Further notice in the figure that the exponential R_i are limited by 4 hours, while the Pareto R_i stretch to as high as 61 hours.

While it is clear that node arrival instants t_v are uncorrelated with lifespans $[a_i, d_i]$ of other nodes, the same observation holds for random points τ_i at which the i-th neighbor of v fails. We extensively experimented with the active model, in which additional node selection occurred at instants τ_i, and found that all R_i obtained in this process also followed (1) very well (not shown for brevity).

3.2 Resilience Analysis

Throughout the paper, we study resilience of P2P systems using two main metrics – the time before all neighbors of v are simultaneously in the failed state and the probability of this occurring before v decides to leave the system. We call the former metric *isolation time* T and the latter *probability of isolation* π. Recall that the passive model follows a simple pure-death degree evolution process illustrated in Figure 1(a). In this environment, a node is considered isolated after its last surviving neighbor fails. Thus, T is equal to the maximum residual lifetime among all neighbors and its expectation can be written as (using the fact that T is a non-negative random variable) [36]:

$$E[T] = \int_0^\infty \left[1 - \frac{1}{E[L_i]^k}\left(\int_0^x (1 - F(z))dz\right)^k\right]dx, \quad (3)$$

which leads to the following two results after straightforward integration.

THEOREM 1. *Assume a passive k-regular graph. Then, for exponential lifetimes:*

$$E[T] = \frac{1}{\lambda}\sum_{i=1}^{k}\binom{k}{i}\frac{(-1)^{i+1}}{i} \quad (4)$$

and for Pareto lifetimes with $\alpha > 2$:

$$E[T] = -\beta\left[1 + \frac{\Gamma\left(\frac{1}{1-\alpha}\right)k!}{(\alpha-1)\Gamma\left(k+2-\frac{\alpha}{\alpha-1}\right)}\right]. \quad (5)$$

Note that the gamma function in the numerator of (5) is negative due to $\alpha > 1$, which explains the $-\beta$ term outside the brackets. Simulation results of (4)-(5) are shown in Figure 4(a) for the average lifetime $E[L_i]$ equal to 0.5 hours. Note that in the figure, simulations are plotted as isolated points and the two models as continuous lines. As the figure shows, simulation results for both exponential and Pareto distributions match the corresponding model very well. We also observe that for the same degree and average lifetime, Pareto nodes exhibit longer average times to isolation. For $k = 10$, $E[T]$ is 1.46 hours given exponential lifetimes and 4.68 hours given Pareto lifetimes. This difference was expected since T is determined by the residual lives of the neighbors, who in the Pareto case have large R_i and stay online longer than newly arriving peers.

We next focus on the probability that isolation occurs within the lifespan of a given user. Consider a node v with lifetime L_v. This node is forced to disconnect from the system only if L_v is greater than T, which happens with probability $\pi = P(T < L_v) = \int_0^\infty F_T(x)f(x)dx$, where $F_T(x) = F_R(x)^k$ is the CDF of time T and $f(x)$ is the PDF of user lifetimes. This leads to:

$$\pi = \frac{1}{E[L_i]^k}\int_0^\infty \left(\int_0^x (1 - F(z))dz\right)^k f(x)dx. \quad (6)$$

Next, we study two distributions $F(x)$ and demonstrate the effect of tail-weight on the local resilience of the system.

THEOREM 2. *Assume a passive k-regular graph. Then, for exponential lifetimes:*

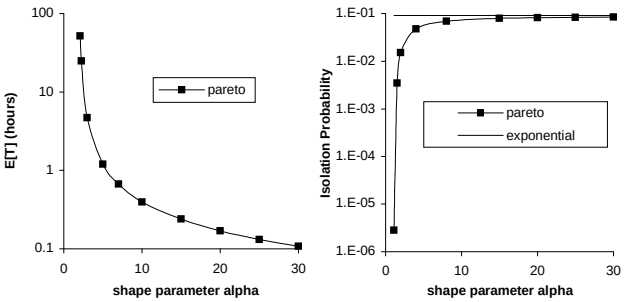
$$\pi = \frac{1}{k+1} \quad (7)$$

and for Pareto lifetimes with $\alpha > 1$:

$$\pi = \frac{\Gamma\left(1+\frac{\alpha}{\alpha-1}\right)k!}{\Gamma\left(k+1+\frac{\alpha}{\alpha-1}\right)}. \quad (8)$$

The exponential part of this lemma was expected from the memoryless property of exponential distributions [28], [36]. Hence, when a new node v joins a P2P system with exponentially distributed lifetimes L_i, it will be forced to disconnect if and only if it can outlive k other random nodes that started at the same time t_v. From Lemma 1, we already know that this happens with probability $1/(k+1)$.

The accuracy of (7)-(8) is shown in Figure 4(b), which plots π obtained in simulations together with that predicted by the models. The simulation again uses a hypothetical P2P system with $n = 1000$ nodes and $E[L_i] = 0.5$. As the figure shows, simulations agree with predicted results well.



(a) $E[T]$ (b) π

Figure 5: Impact of tail weight on the average time to isolation and probability π for $k = 10$ and Pareto lifetimes.

3.3 Discussion

Notice from Figure 4(b) that the Pareto π decays quicker and always stays lower than the exponential π. To better understand the effect of α on the isolation probability in the rather cryptic expression (8), we first show that for all choices of α, Pareto systems are more resilient than exponential. We then show that as $\alpha \to \infty$, (8) approaches from below its upper bound (7).

Setting $c = \Gamma(1+\frac{\alpha}{\alpha-1})$, re-write (8) expanding the gamma function in the denominator:

$$\pi = \frac{ck!}{(k+\frac{\alpha}{\alpha-1})!} \approx c\Big(k+1+\frac{1}{2(\alpha-1)}\Big)^{-\alpha/(\alpha-1)} \qquad (9)$$

and notice that (9) always provides a faster decay to zero as a function of k than (7). For the Pareto example of $\alpha = 3$ shown in Figure 4(b), π follows the curve $(k + 1.25)^{-1.5}$, which decays faster than the exponential model by a factor of $\sqrt{k}$. This difference is even more pronounced for distributions with heavier tails. For example, (8) tends to zero as $(k + 1.5)^{-2}$ for $\alpha = 2$ and as $(k + 6)^{-11}$ for $\alpha = 1.1$. The effect of tail-weight on isolation dynamics is shown in Figure 5 where small values of α indeed provide large $E[T]$ and small π. Figure 5(b) also demonstrates that as shape α becomes large, the Pareto distribution no longer exhibits its "heavy-tailed" advantages and is essentially reduced to the exponential model. This can also be seen in (9), which tends to $1/(k + 1)$ for $\alpha \to \infty$.

Given the above discussion, it becomes apparent that it is possible to make π arbitrarily small with very heavy-tailed distributions (e.g., $\alpha = 1.05$ and $k = 20$ produce $\pi = 3.7 \times 10^{-12}$). While these results may be generally encouraging for networks of non-human devices with controllable characteristics, most current peer-to-peer systems are not likely to be satisfied with the performance of the passive model since selection of α is not possible in the design of a typical P2P network and isolation probabilities in (8) are unacceptably high for $\alpha > 1.5$. The second problem with the passive framework is that its application to real systems requires *accurate* knowledge of the shape parameter α, which may not be available in practice.

We overcome both problems in the next two section, where we show that active node replacement significantly increases resilience and that all Pareto distributions have a reasonably tight upper bound on π that does not depend on α.

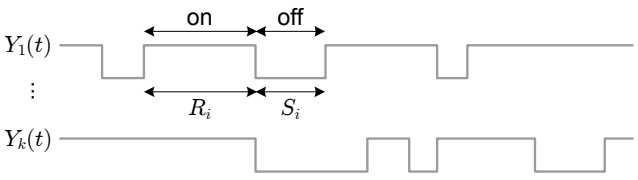


Figure 6: On/off processes $\{Y_i(t)\}_{i=1}^k$ depicting neighbor failure and replacement.

4. ACTIVE LIFETIME MODEL: STATIONARY ANALYSIS

To reduce the rate of isolation and repair broken routes in P2P networks, previous studies have suggested distributed recovery algorithms in which failed neighbors are dynamically replaced with nodes that are still alive. In this section, we offer a model for this strategy, derive the expected value of T using stationary techniques in renewal process theory, and analyze performance gains of this framework compared to the passive case. In the next section, we apply the theory of rare events for mixing processes to $W(t)$ and derive a reasonably good upper bound on π.

It is natural to assume that node failure in P2P networks can be detected through some keep-alive mechanism, which includes periodic probing of each neighbor, retransmission of lost messages, and timeout-based decisions to search for a replacement. We do not dwell on the details of this framework and assume that each peer v is capable of detecting neighbor failure through some transport-layer protocol. The second step after node failure is detected is to repair the "failed" zone of the DHT and restructure certain links to maintain consistency and efficiency of routing (non-DHT systems may utilize a variety of random neighbor-replacement strategies [9], [13], [32]). We are not concerned with the details of this step either and generically combine both failure detection and repair into a random variable called S_i, which is the total "search" time for the i-th replacement in the system.

4.1 Preliminaries

In the active model, each neighbor i ($1 \le i \le k$) of node v is either alive at any time t or its replacement is being sought from among the remaining nodes in the graph. Thus, neighbor i can be considered in the *on* state at time t if it is alive or in the *off* state otherwise. This neighbor failure/replacement procedure can be modeled as an on/off process $Y_i(t)$:

$$Y_i(t) = \begin{cases} 1 & \text{neighbor } i \text{ alive at } t \\ 0 & \text{otherwise} \end{cases}. \qquad (10)$$

This framework is illustrated in Figure 6, which shows the evolution of k neighbor processes $Y_1(t), \ldots, Y_k(t)$. Using this notation, the degree of node v at time t is equal to $W(t) = \sum_{i=1}^{k} Y_i(t)$. Similar to our definition in Section 2.1, a node is isolated at such time T when all of its neighbors are simultaneously in the *off* state (see Figure 1(b)). Thus, the maximum time a node can spend in the system before it is forced to disconnect can be formalized as the *first hitting time* of process $W(t)$ on level 0:

$$T = \inf(t > 0 : W(t) = 0 | W(0) = k). \qquad (11)$$

Notice that under proper selection of the tail-weight of the

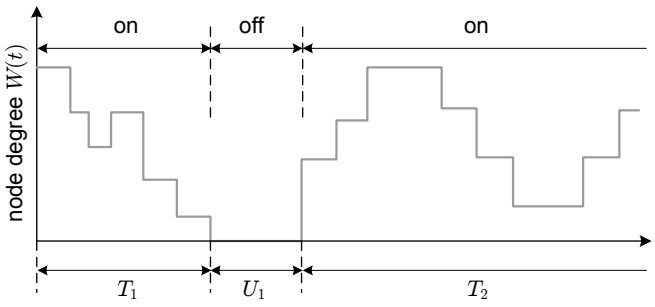


Figure 7: On/off regenerative model of process $W(t)$.

lifetime distribution (i.e., the length of *on* periods), $W(t)$ becomes a super-position of heavy-tailed on/off processes and may exhibit self-similarity for sufficiently large k [15], [18], [35]. Due to limited space, we omit log-log variance plots that confirm this effect, but note that to our knowledge, the fact that node degree in P2P networks may become self-similar has not been documented before.

4.2 Expected Time to Isolation

In what follows in the rest of this section, we apply the theory of regenerative processes to $W(t)$ and derive a closed-form expression for $E[T]$. We start with a simple lemma.

LEMMA 3. *The steady-state probability of finding a neighbor in the on state is given by:*

$$p = \lim_{t\to\infty} P(Y_i(t) = 1) = \frac{E[R_i]}{E[S_i] + E[R_i]}, \qquad (12)$$

where $E[S_i]$ is the mean node-replacement delay and $E[R_i]$ is the expected residual lifetime.

This result immediately leads to the probability of finding stationary process $W(t)$ in any of its $k+1$ states.

COROLLARY 1. *The steady-state distribution of $W(t)$ is binomial with parameters k and p:*

$$\lim_{t\to\infty} P(W(t) = m) = \binom{k}{m} p^m (1-p)^{k-m}. \qquad (13)$$

Both (12)-(13) match simulation results very well (not shown for brevity) and do not depend on the distribution of search delays or user lifetimes. Next, notice that it is convenient to also view $W(t)$ as an alternating on/off process, where each *on* period corresponds to $W(t) > 0$ and each *off* period to $W(t) = 0$. This is illustrated in Figure 7, where U_j is the duration of the j-th *off* cycle and T_j is the length of the j-th *on* cycle. Then we have the following result.

THEOREM 3. *Assuming asymptotically small search delays with $E[S_i] \ll E[R_i]$, the expected time a node can stay in the system before isolation is:*

$$E[T] \approx \frac{E[S_i]}{k}\left[\left(1 + \frac{E[R_i]}{E[S_i]}\right)^k - 1\right], \qquad (14)$$

where $E[S_i]$ is the mean search time and $E[R_i]$ is the expected residual lifetime.

PROOF. Our goal is to determine the expected duration of the first cycle (i.e., $E[T_1]$) shown in Figure 7. The proof consists of two parts: we first argue that the length of cycle T_1 is similar to that of the remaining cycles $T_j, j \geq 2$, and then apply Smith's theorem to $W(t)$ to derive $E[T_j], j \geq 2$.

First, notice that cycle T_1 is different from the other *on* periods since it always starts from $W(t) = k$, while the other *on* cycles start from $W(t) = 1$. However, since we already assumed that the search times are sufficiently small, $W(t)$ at the beginning of each *on* period almost immediately "shoots back" to $W(t) = k$. This can be shown using arguments from large-deviations theory [34], which derives bounds on the return time of the system from very rare states back to its "most likely" state. This generally makes cycles T_1 and T_j ($j \geq 2$) different by a value that is negligible compared to $E[T]$ in real-life situations (see examples after the proof).

We next derive the expected length of $T_j, j \geq 2$. Approximating points τ_j when $W(t)$ goes into the j-th *off* state (i.e., makes a transition from 1 to 0) as regenerative instances and applying Smith's theorem to $W(t)$ [28], the probability of finding the process in the *off* state at any random time t is given by:

$$\lim_{t\to\infty} P(W(t) = 0) \approx \frac{E[U_j]}{E[T_j] + E[U_j]}. \qquad (15)$$

Notice that (15) can also be expressed from (13):

$$\lim_{t\to\infty} P(W(t) = 0) = \left(\frac{E[S_i]}{E[S_i] + E[R_i]}\right)^k. \qquad (16)$$

Equating (15) and (16) and solving for $E[T_j]$, we get:

$$E[T_j] \approx E[U_j]\left[\left(\frac{E[R_i] + E[S_i]}{E[S_i]}\right)^k - 1\right]. \qquad (17)$$

Next, we compute $E[U_j]$. As before, suppose that the first instant of the j-th *off* cycle of $W(t)$ starts at time τ_j. At this time, there are $k-1$ already-failed neighbors still "searching" for their replacement and one neighbor that just failed at time τ_j. Thus, U_j is the minimum time needed to find a replacement for the last neighbor or for one of the on-going searches to complete.

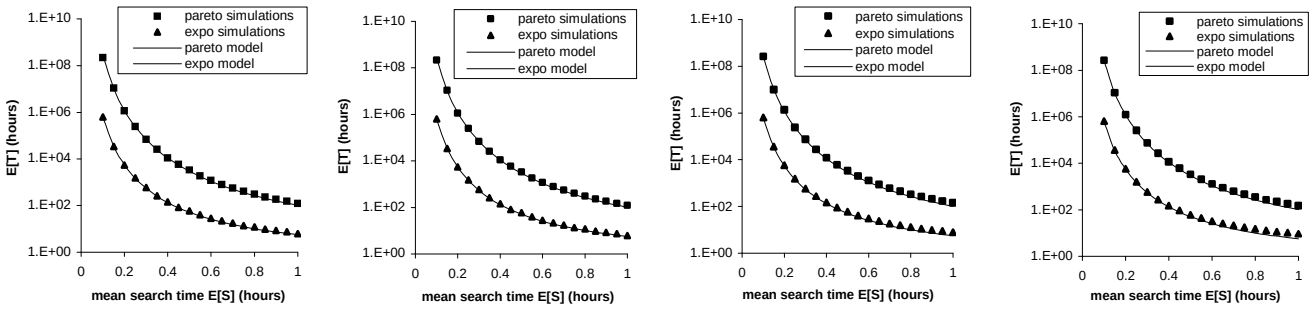
More formally, suppose that $V_1, \ldots, V_{k-1}$ represent the remaining replacement delays of the $k-1$ already-failed neighbors and S_k is the replacement time of the last neighbor.[2] Then duration U_j of the current *off* cycle is $U_j = \min\{V_1, \ldots, V_{k-1}, S_k\}$. Assuming that $F_S(x)$ is the CDF of search times S_i, the distribution of $V = \min\{V_1, \ldots, V_{k-1}\}$ is given by [36]:

$$F_V(x) = 1 - \left(1 - \frac{1}{E[S_i]}\int_0^x (1 - F_S(z))dz\right)^{k-1}. \qquad (18)$$

Notice that U_j can also be written as $U_j = \min\{V, S_k\}$ and its expectation is:

$$E[U_j] = \int_0^\infty (1 - F_S(x))(1 - F_V(x))\,dx. \qquad (19)$$

[2]Strictly speaking, $V_1, \ldots, V_{k-1}$ may have different distributions that depend on the difference between the time of each neighbor's failure and τ_j; however, for very small $E[S_i]$, this distinction does not have a noticeable impact on (14).



(a) uniform S_i (b) binomial S_i (c) exponential S_i (d) Pareto S_i with $\alpha = 3$

Figure 8: Comparison of model (14) to simulation results with $E[L_i] = 0.5$ and $k = 10$.

Substituting (18) into (19), we get:

$$E[U_j] = \int_0^\infty (1 - F_S(x)) \left(1 - \frac{\int_0^x (1 - F_S(z))dz}{E[S_i]}\right)^{k-1} dx. \tag{20}$$

Setting $y = \int_0^x (1 - F_S(z))dz$ in (20), we have:

$$E[U_j] = \int_0^{E[S_i]} \left(1 - \frac{y}{E[S_i]}\right)^{k-1} dy = \frac{E[S_i]}{k}, \tag{21}$$

which leads to (14). □

Note that for small search delays this result does not generally depend on the distribution of residual lifetimes R_i or search times S_i, but only on their expected values. Figure 8 shows $E[T]$ obtained in simulations in a system with 1000 nodes, $k = 10$, and four different distributions of search delay. In each part (a)-(d) of the figure, the two curves correspond to exponential and Pareto lifetimes with mean 30 minutes (as before, the models are plotted as solid lines and simulations are drawn using isolated points). Notice in all four subfigures that the model tracks simulation results for over 7 orders of magnitude and that the expected isolation time is in fact not sensitive to the distribution of S_i.

As in the passive model, the exponential distribution of lifetimes provides a lower bound on the performance of any Pareto system (since exponential $E[R_i]$ is always smaller than the corresponding Pareto $E[R_i]$). Further observe that the main factor that determines $E[T]$ is the *ratio* of $E[R_i]$ to $E[S_i]$ and not their individual values. Using this insight and Figure 8, we can conclude that in systems with 10 neighbors and expected search delay at least 5 times smaller than the mean lifetime, $E[T]$ is at least *one million* times larger than the mean session length of an average user. Furthermore, this result holds for exponential as well as Pareto distributions with arbitrary α. This is a significant improvement over the results of the passive model in Section 3.

4.3 Chord Example

We now phrase the above framework in more practical terms and study the resilience of Chord as a function of node degree k. For the sake of this example, suppose that each node relies on a keep-alive protocol with timeout δ. Then, the distribution of failure-detection delays is uniform in $[0, \delta]$ depending on when the neighbor died with respect to the nearest ping message. The actual search time to find a replacement may be determined by the average number of application-layer hops between each pair of users and the average Internet delay d between the overlay nodes in the system. Using the notation above, we have the following re-statement of the previous theorem.

Timeout δ	$k = 20$	$k = 10$	$k = 5$
20 sec	10^{41} years	10^{17} years	188,034 years
2 min	10^{28} years	10^{11} years	282 years
45 min	404,779 years	680 days	49 hours

Table 1: Expected time $E[T]$ for $E[R_i] = 1$ hour.

COROLLARY 2. *Assuming that δ is the keep-alive timeout and d is the average Internet delay between the P2P nodes, the expected isolation time in Chord is given by:*

$$E[T] = \frac{\delta + d\log_2 n}{2k} \left(1 + \frac{2E[R_i]}{\delta + d\log_2 n}\right)^k. \tag{22}$$

Consider a Chord system with the average inter-peer delay $d = 200$ ms, $n = 1$ million nodes (average distance 10 hops), and $E[R_i] = 1$ hour. Table 1 shows the expected time to isolation for several values of timeout δ and degree k. For small keep-alive delays (2 minutes or less), even $k = 5$ provides longer expected times to isolation than the lifetime of any human being. Also notice that for $\delta = 2$ minutes, Chord's default degree $k = 20$ provides more years before expected isolation than there are molecules in a glass of water [2].

Since routing delay d in the overlay network is generally much smaller than keep-alive timeout δ, the diameter of the graph does not usually contribute to the resilience of the system. In other cases when $d \log n$ is comparable to δ, P2P graphs with smaller diameter may exhibit higher resilience as can be observed in (22).

4.4 Real P2P Networks

Finally, we address the practicality of the examples shown in the paper so far. It may appear that $E[R_i] = 1$ hour is rather large for current P2P systems since common experience suggests that many users leave within several minutes of their arrival into the system. This is consistent with our Pareto model in which the majority of users have very small online lifetimes, while a handful of users that stay connected for weeks contribute to the long tails of the distribution. For Pareto lifetimes with $\alpha = 3$ and $E[R_i] = 1$ hour, the mean

Timeout δ	$k = 10$	$k = 5$	$k = 2$
20 sec	10^{29} years	10^{11} years	4.7 years
2 min	10^{23} years	10^{8} years	336 days
45 min	10^{11} years	1,619 years	16 days

Table 2: Expected time $E[T]$ for Pareto lifetimes with $\alpha = 2.06$ ($E[L_i] = 0.93$ hours, $E[R_i] = 16.6$ hours).

online stay is only 30 minutes and 25% of the users depart within 6 minutes of their arrival. In fact, this rate of turnaround is quite aggressive and exceeds that observed in real P2P systems by a factor of two [31].

We should also address the results shown in [8], which suggest that the empirical distribution of user lifetimes in real P2P networks follows a Pareto distribution with shape parameter $\alpha = 1.06$. Such heavy-tailed distributions result in $E[R_i] = E[T] = \infty$ and do not lead to much interesting discussion. At the same time, notice that while it is hypothetically possible to construct a P2P system with $\alpha = 1.06$, it can also be argued that the measurement study in [8] sampled the *residuals* rather than the actual lifetimes of the users. This is a consequence of the "snapshots" taken every 20 minutes, which missed all peers with $L_i < 20$ minutes and shortened the lifespans of the remaining users by random amounts of time. As such, these results point toward $\alpha = 2.06$, which is a much more realistic shape parameter even though it still produces enormous $E[T]$ for all feasible values of $E[S_i]$. This is demonstrated for Chord's model (22) in Table 2 where the expected lifetime of each user is only double that in Table 1, but $E[T]$ is 5-12 orders of magnitude larger. This is a result of $E[R_i]$ rising from 1 hour in the former case to 16.6 hours in the latter scenario.

5. ACTIVE LIFETIME MODEL: TRANSIENT ANALYSIS

Given the examples in the previous section, it may at first appear that π must automatically be very small since $E[T]$ is so "huge" under all practical conditions. However, in principle, there is a possibility that a large mass of T is concentrated on very small values and that a handful of extremely large values skew the mean of T to its present location. We additionally are interested in more than just knowing that π is "small" – we specifically aim to understand the order of this value for different $E[S_i]$.

As in previous sections, let L_v denote the lifetime of v and T the random time before v's neighbors force an isolation. Notice that $\pi = P(T < L_v) = \int_0^\infty F_T(t)f(t)dt$ is an integral of the CDF function $F_T(t) = P(T < t)$ of the first hitting time of process $W(t)$ on level 0. The exact distribution of T is difficult to develop in closed-form since it depends on *transient* properties of a complex process $W(t)$. To tackle this problem, we first study the asymptotic case of $E[S_i] \ll E[R_i]$ and apply results from the theory of rare events for Markov jump processes [3], [34] to derive a very accurate formula for π assuming exponential lifetimes. We then use this result to upper-bound the Pareto version of this metric.

5.1 Exponential Lifetimes

We start with exponential lifetimes and assume reasonably small search times. For $E[S_i]$ larger than $E[R_i]$, accurate isolation probabilities are available from the passive model in Section 3.

THEOREM 4. *For exponential lifetimes L_i and asymptotically small search delays with $E[S_i] \ll E[R_i]$, the probability of isolation converges to:*

$$\pi \approx \frac{E[L_i]}{E[T]}. \tag{23}$$

PROOF. We again proceed in several steps. We first assume exponential search times and construct a Markov chain based on $W(t)$. We then bind $F_T(t) = P(T < t)$ using inequalities for rare events in Markov chains and complete the proof by extending this result to asymptotically small non-exponential search times.

Given exponential S_i, notice that $W(t)$ can be viewed as a continuous-time Markov chain, where the time spent in each state j before making a transition to state $j-1$ is the minimum of exactly j exponential variables (i.e., the time to the next failure). Assume that the CDF of R_i is $F_R(x) = 1 - e^{-\lambda x}$, where $\lambda = 1/E[L_i]$. Then the CDF of $\min\{R_1, \ldots, R_j\}$ is $1 - (1 - F_R(x))^j = 1 - e^{-\lambda j x}$, which is another exponential variable with rate λj. Next notice that the delays before $W(t)$ makes a transition from state j to $j+1$ (i.e., upon recovering a neighbor) are given by the minimum of $k-j$ residual search times, which is yet another exponential random variable with rate $(k-j)\mu$, where $\mu = 1/E[S_i]$.

To bound the CDF of T, one approach is to utilize classical analysis from Markov chains that relies on numerical exponentiation of transition (or rate) matrices; however, it does not lead to a closed-form solution for $P(T < L_v)$. Instead, we apply a result for rare events in Markov chains due to Aldous *et al.* [3], which shows that T asymptotically behaves as an exponential random variable with mean $E[T]$:

$$|P(T > t) - e^{-t/E[T]}| \leq \frac{\tau}{E[T]}, \tag{24}$$

where $E[T]$ is the expected time between the visits to the rare state 0 and τ is the relaxation time of the chain. Rewriting (24) in terms of $F_T(t) = P(T < t)$ and applying Taylor expansion to $e^{-t/E[T]}$:

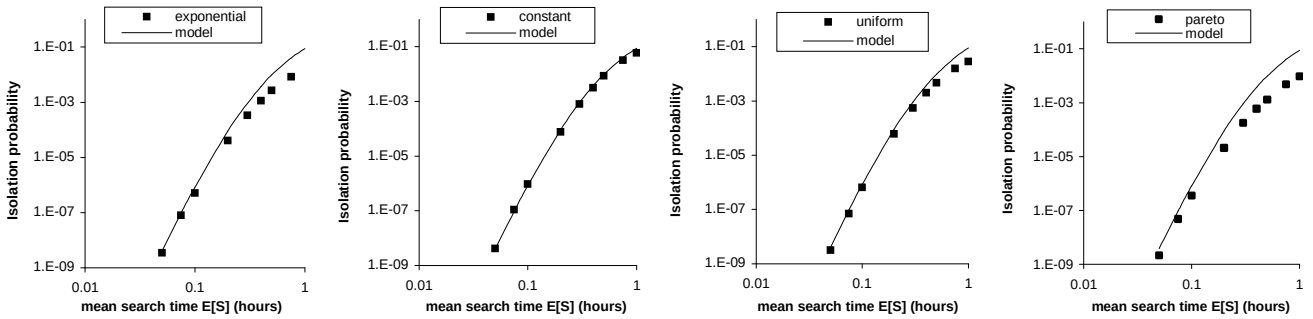
$$\frac{t-\tau}{E[T]} \leq F_T(t) \leq \frac{t+\tau}{E[T]}. \tag{25}$$

Next, recall that relaxation time τ is the inverse of the second largest eigenvalue of $-Q$, where Q is the rate matrix of the chain. For birth-death chains, matrix Q is tri-diagonal with $Q(i,i) = -\sum_{j \neq i} Q(i,j)$:

$$Q = \begin{bmatrix} -k\lambda & k\lambda & \cdots & 0 \\ \mu & -\mu-(k-1)\lambda & (k-1)\lambda & \\ 0 & \cdots & \cdots & \lambda \\ 0 & \cdots & k\mu & -k\mu \end{bmatrix}. \tag{26}$$

We treat state $W(t) = 0$ as non-absorbing and allow the chain to return back to state 1 at the rate $k\mu$. Then, the second largest eigenvalue of this matrix is available in closed-form (e.g., [19]) and equals the sum of individual rates: $\lambda_2 = 1/\tau = \lambda + \mu$. Noticing that:

$$\tau = \frac{1}{\lambda+\mu} = \frac{1}{1/E[L_i] + 1/E[S_i]} \approx E[S_i], \tag{27}$$



(a) exponential S_i (b) constant S_i (c) uniform S_i (d) Pareto S_i with $\alpha = 3$

Figure 9: Comparison of model (23) to simulation results for exponential lifetimes with $E[L_i] = 0.5$ and $k = 10$.

we conclude that τ is on the order of $E[S_i]$ and is generally very small. Writing $\tau \approx E[S_i]$ and integrating the upper bound of (25) over all possible values of lifetime t, we get:

$$\pi \leq \int_0^\infty \frac{(t+\tau)f(t)dt}{E[T]} = \frac{E[L_i] + E[S_i]}{E[T]}. \qquad (28)$$

We similarly obtain a lower bound on π, which is equal to $(E[L_i] - E[S_i])/E[T]$. Neglecting small $E[S_i]$, observe that both bounds reduce to (23).

Our final remark is that for non-exponential, but asymptotically small search delays, $W(t)$ can usually be approximated by an equivalent, but quickly-mixing process and that bounds similar to (25) are reasonably accurate regardless of the distribution of S_i [1]. □

Simulation results of π are shown in Figure 9 using four distributions of search time – exponential with rate $\lambda = 1/E[S_i]$, constant equal to $E[S_i]$, uniform in $[0, 2E[S_i]]$, and Pareto with $\alpha = 3$. As shown in the figure, all four cases converge with acceptable accuracy to the asymptotic formula (23) and achieve isolation probability $\pi \approx 3.8 \times 10^{-9}$ when the expected search time reduces to 3 minutes. Also note that for large $E[S_i]$, model (23) provides an *upper* bound on the actual π for all four cases.

5.2 Heavy-Tailed Lifetimes

Although it would be nice to obtain a similar result $\pi \approx E[L_i]/E[T]$ for the Pareto case, unfortunately the situation with a superposition of heavy-tailed on/off processes is different since $W(t)$ is slowly mixing and the same bounds no longer apply. Intuitively, it is clear that large values of $E[T]$ in the Pareto case are caused by a handful of users with enormous isolation delays, while the majority of remaining peers acquire neighbors with short lifetimes and suffer isolation almost as quickly as in the exponential case. Consider an example that illustrates this effect and shows that huge values of $E[T]$ in Pareto systems have little impact on π. For 10 neighbors, $\alpha = 3$ and $\lambda = 2$ ($E[L_i] = 30$ minutes), and constant search time $s = 6$ minutes, the Pareto $E[T]$ is larger than the exponential $E[T]$ by a factor of 865. However, the ratio of their isolation probabilities is only 5.7. For $\alpha = 2.5$ and $\lambda = 1.5$ ($E[L_i] = 40$ minutes), the expected times to isolation differ by a factor of 8.1×10^5, but the ratio of their π is only 7.5.

It may be possible to derive an accurate approximation for Pareto π; however, one may also argue that the usefulness of such a result is limited given that shape parameter α and the distribution of user lifetimes (lognormal, Pareto, etc.) are often not known accurately. We leave the exploration of this problem for future work and instead utilize the exponential metric (23) as an upper bound on π in systems with sufficiently heavy-tailed lifetime distributions. The result below follows from the fact that heavy-tailed L_i imply stochastically larger residual lifetimes R_i and a straightforward expansion of $E[T]$ in (23).

π	Uniform $p = 1/2$	Lifetime P2P	Mean search time $E[S_i]$		
			6 min	2 min	20 sec
10^{-6}	20	Bound (29)	10	7	5
		Simulations	9	6	4
10^{-9}	30	Bound (29)	14	9	6
		Simulations	13	8	6
10^{-12}	40	Bound (29)	18	12	8
		Simulations	17	11	7

Table 3: Minimum degree needed to achieve a certain π for Pareto lifetimes with $\alpha = 2.06$ and $E[L_i] = 0.5$ hours.

COROLLARY 3. *For an arbitrary distribution of search delays and any lifetime distribution $F(x)$ with an exponential or heavier tail, which includes Pareto, lognormal, Weibull, and Cauchy distributions, the following upper bound holds:*

$$\pi \leq \frac{kE[L_i]E[S_i]^{k-1}}{(E[L_i] + E[S_i])^k - E[S_i]^k} \approx \frac{kE[L_i]E[S_i]^{k-1}}{(E[L_i] + E[S_i])^k}. \qquad (29)$$

For example, using 30-minute average lifetimes, 9 neighbors per node, and 1-minute average node replacement delay, the upper bound in (29) equals 1.02×10^{-11}, which allows the joining users in a 100-billion node network to stay connected to the graph for their entire lifespans with probability $1 - 1/n$. Using the uniform failure model of prior work and $p = 1/2$, each user required 37 neighbors to achieve the same π *regardless of the actual dynamics of the system.*

Even though exponential π is often several times larger than the Pareto π (the exact ratio depends on shape α), it turns out that the difference in node degree needed to achieve a certain level of resilience is usually negligible. To illustrate this result, Table 3 shows the minimum degree k that ensures a given π for different values of search time $E[S_i]$ and Pareto lifetimes with $\alpha = 2.06$ (to maintain the mean lifetime 30 minutes, the distribution is scaled using

$\beta = 0.53$). The column "uniform $p = 1/2$" contains degree k that can be deduced from the p–percent failure model (for $p = 1/2$) discussed in previous studies [33]. Observe in the table that the exponential case in fact provides a tight upper bound on the actual minimum degree and that the difference between the two cases is at most 1 neighbor.

5.3 Irregular Graphs

The final issue addressed in this section is whether P2P networks can become more resilient if node degree is allowed to vary from node to node. It is sometimes argued [10], [31] that graphs with a heavy-tailed degree distribution exhibit highly resilient characteristics and are robust to node failure. Another question raised in the literature is whether DHTs are more resilient than their unstructured counterparts such as Gnutella. In this section, we prove that, given the assumptions used so far in the paper, k-regular graphs offer the highest local resilience among all systems with a given average degree. This translates into "optimality" of DHTs as long as they can balance their zone-sizes and distribute degree evenly among the peers.

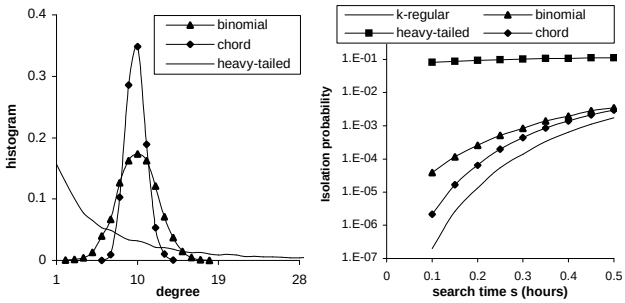
Consider a P2P system in which node degrees $k_1, \ldots, k_n$ are drawn from an arbitrary distribution with mean $E[k_i]$. Using Jensen's inequality for convex functions and the upper bound in (29), the following result follows immediately.

THEOREM 5. *Assuming that lifetimes are independent of node degree and are not used in the neighbor-selection process, regular graphs are the most resilient for a given average degree $E[k_i]$.*

To demonstrate the effect of node degree on isolation probability in irregular graphs, we examine three systems with 1000 nodes: 1) Chord with a random distribution of out-degree, which is a consequence of imbalance in zone sizes; 2) a $G(n,p)$ graph with binomial degree for $p = 0.5$; and 3) a heavy-tailed graph with Pareto degree for $\alpha = 2.5$ and $\beta = 15$. We selected these parameters so that each of the graphs had a mean degree $E[k_i]$ equal to 10. The distribution of degree in these graphs is shown in Figure 10(a). Notice that Chord has the lowest variance and its probability mass concentration around the mean is the best of the three systems. The binomial case is slightly worse, while the heavy-tailed graph is the worst. According to Theorem 5, all of these systems should have larger isolation probabilities than those of 10-regular graphs and should exhibit performance inverse proportional to the variance of their degree.

Simulation results of π are shown in Figure 10(b) for Pareto lifetimes with $\alpha = 3$ and $E[L_i] = 0.5$ hours (search times are constant). Observe in the figure that the k-regular system is in fact better than the irregular graphs and that the performance of the latter deteriorates as $Var[k_i]$ increases. For $s = 0.1$ (6 minutes), the k-regular graph offers π lower than Chord's by a factor of 10 and lower than that in $G(n,p)$ by a factor of 190. Furthermore, the heavy-tailed P2P system in the figure exhibits the same poor performance regardless of the search time s and allows users to become isolated $10^2 - 10^6$ times more frequently than in the optimal case, all of which is caused by 37% of the users having degree 3 or less.

Thus, in cases when degree is independent of user lifetimes, we find no evidence to suggest that unstructured P2P systems with a heavy-tailed (or otherwise irregular) degree can provide better resilience than k-regular DHTs.



(a) degree histogram (b) π

Figure 10: (a) Degree distribution in irregular graphs. (b) User isolation probability π of irregular graphs (average degree $E[k_i] = 10$, lifetimes are Pareto with $E[L_i] = 0.5$ hours).

6. GLOBAL RESILIENCE

We finish the paper by analyzing the probability of network partitioning under uniform node failure and showing that this metric has a simple expression for a certain family of graphs, which includes many proposed P2P networks. We then apply this insight to lifetime-based systems and utilize the earlier derived metric π to characterize the evolution of P2P networks under churn.

6.1 Classical Result

One may wonder how local resilience (i.e., absence of isolated vertices) of P2P graphs translates into their global resilience (i.e., connectivity of the entire network). While this topic has not received much attention in the P2P community, it has been extensively researched in random graph theory and interconnection networks. Existing results for classical random graphs have roots in the work of Erdös and Rényi in the 1960s and demonstrate that *almost every* (i.e., with probability $1 - o(1)$ as $n \to \infty$) random graph including $G(n,p)$, $G(n,M)$, and $G(n,k_{out})$ is connected if and only if it has no isolated vertices [6], i.e.,

$$P(G \text{ is connected}) = P(X = 0) \quad \text{as } n \to \infty, \tag{30}$$

where X is the number of isolated nodes after the failure. After some manipulation, this result can be translated to apply to unstructured P2P networks, where each joining user draws some number of random out-degree neighbors from among the existing nodes (see below for simulations that confirm this).

For deterministic networks, connectivity of a graph G after node/edge failure has also received a fair amount of attention (e.g., [7], [21]). In interconnection networks, exact formulas for the connectivity of deterministic graphs exist [21]; however, they require computation of NP-complete metrics and no closed-form solution is available even for the basic hypercube. However, from the perspective of random graph theory, it has been shown [6], [7] that hypercubes with faulty elements asymptotically behave as random graphs and thus almost surely disconnect with isolated nodes as n becomes large.

Even though the necessary condition for a deterministic G to satisfy (30) is unknown at this time, sufficient conditions can be extrapolated from the proofs of this relationship for the hypercube [6]. The general requirement on G is that its

p	$P(G$ is weakly connected)			$P(G$ has no isolated nodes)			$P(G$ disconnects with isolated nodes)		
	Chord	Symphony	Gnutella	Chord	Symphony	Gnutella	Chord	Symphony	Gnutella
0.5	0.99996	0.99768	0.86257	0.99996	0.99768	0.86260	1	1	0.999782
0.55	0.99918	0.98750	0.58042	0.99918	0.98750	0.58064	1	1	0.999476
0.6	0.99354	0.93914	0.17081	0.99354	0.93917	0.17148	1	0.99958	0.999192
0.65	0.95001	0.75520	0.00547	0.95004	0.75527	0.00560	0.99946	0.99971	0.999869
0.7	0.72619	0.31153	0	0.72650	0.31205	0	0.99980	0.99922	1
0.8	0.00040	0	0	0.00043	0	0	0.99997	1	1

Table 4: Global resilience for $n = 16,384$ and out-degree $k = 14$.

expansion (strength of the various cuts) must be no worse than that of the hypercube.[3] Due to limited space and wide variety of deterministic P2P constructs, we do not provide a rigorous re-derivation of this fact, but instead note that Chord [33], Pastry [29], and CAN with the number of dimensions $d = \Theta(\log n)$ [27] can be directly reduced to hypercubes; de Bruijn graphs [16] exhibit better connectivity than hypercubes [23]; and hybrid networks (such as Symphony [24], Randomized Chord [24], and [4]) generally have connectivity no worse than $G(n, k_{out})$.

We next show simulation results that confirm the application of classical result (30) to two types of DHTs and one unstructured Gnutella-like network (all three are directed graphs). Table 4 shows simulations of Chord, Symphony, and a k-regular Gnutella network under uniform p–percent node failure using 100,000 failure patterns. Note that for directed graphs, (30) applies only to the definition of *weak* connectivity, which means that G is disconnected if and only if its *undirected* version G' is and a node $v \in G$ is isolated if and only if it is isolated in G'. Since the total degree (which is the sum of out-degree and in-degree) at each node is different between the graphs in the table, their respective disconnection probabilities are different.

As the table shows, (30) holds with high accuracy and a vast majority of disconnections contain at least one isolated node (the last column of the table). Additional simulations show that (30) applies to Pastry, CAN, de Bruijn graphs, Randomized Chord, and degree-irregular Gnutella. We omit these results for brevity.

6.2 Lifetime-Based Extension

It is not difficult to see that (30) holds for *lifetime-based* P2P graphs and that dynamic P2P networks are also much more likely to develop disconnections around single nodes rather than along boundaries of larger sets S. However, instead of having a single node-failure metric p, we have a probability of isolation π associated with each joining user i. Thus, one may ask a question *what is the probability that the system survives N user joins and stays connected the entire time?* The answer is very simple: assuming Y is a geometric random variable measuring the number of user joins before the first disconnection of the network, we have for almost every sufficiently large graph:

$$P(Y > N) = (1 - \pi)^N. \quad (31)$$

Simulation results of (31) are shown in Table 5 using $N = 1$ million joins and 10,000 iterations per search time,

[3]For each set S in the original graph G, its node boundary must satisfy a certain inequality that is an increasing function of $|S|$ [7]. Graphs that do not fulfill this requirement include trees, cycles, and other weakly connected structures.

Fixed search time (min)	Actual $P(Y > N)$	Model (31)	$q(G)$	$r(G)$
6	0.9732	0.9728	1	1
7.5	0.8118	0.8124	1	1
8.5	0.5669	0.5659	1	1
9	0.4065	0.4028	1	1
9.5	0.2613	0.2645	1	1
10.5	0.0482	0.0471	1	1

Table 5: Comparison of $P(Y > 10^6)$ in k-regular CAN (exponential lifetimes with mean 30 minutes) to model (31). The graph has $d = 6$ dimensions, degree $k = 12$, and $n = 4096$ nodes.

where metric $q(G) = P(X > 0|G$ is disconnected) is the probability that the graph partitions with at least one isolated node and $r(G)$ is the probability that the largest connected component after the disconnection contains exactly $n - 1$ nodes. As the table shows, simulations match the model very well and also confirm that the most likely disconnection pattern of lifetime-based systems includes at least one isolated node (i.e., $q(G) = 1$). In fact, the table shows an even stronger result – for reasonably small search delays, network partitioning almost surely affects only *one* node in the system (i.e., $r(G) = 1$). The same conclusion holds for other P2P graphs, Pareto lifetimes, and random search delays. We omit these results for brevity.

Model (31) suggests that when search delays become very small, the system may evolve for many months or years before disconnection. Consider a 12-regular CAN system with 1-minute search delays and 30-minute average lifetimes. Assuming that $n = 10^6$ and each user diligently joins the system once per day, the probability that the network can evolve for 2,700 years ($N = 10^{13}$ joins) before disconnecting for the first time is 0.9956. The mean delay before the first disconnection is $E[Y] = 1/\pi$ user joins, or 5.9 million years.

7. FUTURE WORK

The discussion at the end of Section 5 suggests that irregular graphs cannot be beneficial, unless *node degree is correlated with user lifetimes*. Notice that users v who stay online longer have a larger probability of disconnection $\pi = P(T < L_v) \approx L_v/E[T]$ (see the proof of Theorem 4). Thus, to achieve resilience higher than that of k-regular graphs, one must assign smaller degree to nodes that have smaller lifetimes, and vice versa. To better understand this concept, which we call *Dynamic Degree Scaling* (DDS), notice that peers with very small L_i do not require 20 neighbors to stay connected for their entire lifespans. In fact, these

nodes often leave even before their first neighbor fails. As the age of a peer increases, it becomes more likely to outlive its original neighbors and suffer an isolation, which clearly warrants an increase in its degree over time. In future work, we plan to understand how DDS allows unstructured P2P systems to evolve into graphs where users that stay online longer are responsible for larger parts of the graph and examine whether these evolution models can be implemented in practice and/or combined with other methods, in which degree depends on unequal peer characteristics (e.g., [9]).

Finally, note that while DDS allows degree k_i to depend on users' lifetime L_i, the actual links between the nodes can be formed randomly and without the knowledge of the other peers' lifetimes or their current age. An alternative strategy is to allow each node to prefer connections to peers that have the largest age metric at current time t [8], which requires a different model and possibly has its own optimal strategy. We plan to investigate this direction in future work as well.

8. CONCLUSION

This paper examined two aspects of resilience in dynamic P2P systems – ability of each user to stay connected to the system in the presence of frequent node departure and partitioning behavior of the network as $n \to \infty$. We found that under all practical search times, k-regular graphs were much more resilient than traditionally implied [16], [22], [33] and further showed that dynamic P2P networks could almost surely remain connected as long as no user suffered simultaneous neighbor failure. We also demonstrated that varying node degree from peer to peer can have a positive impact on resilience *only* when such decisions are correlated with the users' lifetimes.

9. REFERENCES

[1] M. Abadi and A. Galves, "Inequalities for the Occurrence Times of Rare Events in Mixing Processes. The State of the Art," *Markov Proc. Relat. Fields*, vol. 7, no. 1, 2001.

[2] P.W. Atkins. *Physical Chemistry*. Freeman, NY, 1986.

[3] D.J. Aldous and M. Brown, "Inequalities for Rare Events in Time-Reversible Markov Chains II," *Stochastic Processes and their Applications*, vol. 44, 1993.

[4] J. Aspnes, Z. Diamadi, and G. Shah, "Fault Tolerant Routing in Peer to Peer Systems," *ACM PODC*, 2002.

[5] R. Bhagwan, S. Savage, and G.M. Voelker, "Understanding Availability," *IPTPS*, 2003.

[6] B. Bollobás. *Random Graphs*. Cambridge U. Press, 2001.

[7] Yu.D. Burtin, "Connection Probability of a Random Subgraph of an n-Dimensional Cube," *Probl. Pered. Inf.*, vol. 13, no. 2, April-June 1977.

[8] F.E. Bustemante and Y. Qiao, "Friendships that Last: Peer Lifespan and its Role in P2P Protocols," *Intl. Workshop on Web Caching and Distribution*, September 2003.

[9] Y. Chawathe, S. Ratnasamy, L. Breslau, N. Lanham, and S. Shenker, "Making Gnutella Like P2P Systems Scalable", *ACM SIGCOMM*, 2003.

[10] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin, "Resilience of the Internet to Random Breakdowns," *Physical Review Letters*, vol. 85, no. 21, November 2000.

[11] A. Fiat and J. Saia, "Censorship Resistant Peer-to-Peer Content Addressable Networks," *ACM-SIAM SODA*, 2002.

[12] A. Ganesh and L. Massoulie, "Failure Resilience in Balanced Overlay Networks," *Allerton Conference on Communication, Control and Computing*, 2003.

[13] C. Gkantsidis, M. Mihail, A. Saberi, "Random Walks in Peer-to-Peer Networks," *IEEE INFOCOM*, March 2004.

[14] K. Gummadi, R. Gummadi, S. Gribble, S. Ratnasamy, S. Shenker, and I. Stoica, "The Impact of DHT Routing Geometry on Resilience and Proximity," *ACM SIGCOMM*, 2003.

[15] D. Heath, S. Resnick, and G. Samorodnitsky, "Patterns Of Buffer Overflow In A Class Of Queues With Long Memory In The Input Stream," *Annals of Probability*, 1997.

[16] M.F. Kaashoek and D. Karger, "Koorde: A Simple Degree-optimal Distributed Hash Table, " *IPTPS*, 2003.

[17] F.T. Leighton, B.M. Maggs, and R.K. Sitamaran, "On the Fault Tolerance of Some Popular Bounded-Degree Networks," *IEEE FOCS*, 1995.

[18] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, "On the Self-Similar Nature of Ethernet Traffic," *ACM SIGCOMM*, 1993.

[19] R.B. Lenin and P.R. Parthasarathy, "Transient Analysis in Discrete Time of Markovian Queues with Quadratic Rates," *Southwest J. Pure and Appl. Math.*, July 2000.

[20] D. Leonard, V. Rai, and D. Loguinov, "On Lifetime-Based Node Failure and Resilience of Decentralized Peer-to-Peer Networks (extended version)," *Texas A&M Technical Report*, April 2005.

[21] S. Liu, K-H. Cheng, and X. Liu, "Network Reliability with Node Failures," *Networks*, vol. 35, no. 2, March 2000.

[22] D. Liben-Nowell, H. Balakrishnan, and D. Karger, "Analysis of the Evolution of Peer-to-Peer Systems," *ACM PODC*, 2002.

[23] D. Loguinov, A. Kumar, V. Rai, and S. Ganesh, "Graph-Theoretic Analysis of Peer-to-Peer Systems: Routing Distances and Fault Resilience," *ACM SIGCOMM*, 2003.

[24] G.S. Manku, M. Naor, and U. Weider, "Know thy Neighbor's Neighbor: the Power of Lookahead in Randomized P2P Networks," *ACM STOC*, June 2004.

[25] L. Massoulié, A.-M. Kermarrec, and A. Ganesh, "Network Awareness and Failure Resilience in Self-Organising Overlay Networks," *IEEE Symposium on Reliable and Distributed Systems*, 2003.

[26] G. Pandurangan, P. Raghavan, E. Upfal, "Building Low-Diameter Peer-to-Peer Networks, " *IEEE JSAC*, vol. 21, no. 6, August 2003.

[27] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A Scalable Content-Addressable Network," *ACM SIGCOMM*, 2001.

[28] S. Resnick. *Adventures in Stochastic Processes*. Birkhäuser, Boston, 2002.

[29] A. Rowstron and P. Druschel, "Pastry: Scalable, Decentralized, Object Location and Routing for Large-Scale Peer-to-Peer Systems," *IFIP/ACM Middleware*, 2001.

[30] J. Saia, A. Fiat, S. Gribble, A.R. Karlin, and S. Saroiu, "Dynamically Fault-Tolerant Content Addressable Networks," *IPTPS*, 2002.

[31] S. Saroiu, P.K. Gummadi, and S.D. Gribble, "A Measurement Study of Peer-to-Peer File Sharing Systems," *MMCN*, 2002.

[32] K. Sripanidkulchai, B. Maggs, H. Zhang, "Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems," *IEEE INFOCOM*, March 2003.

[33] I. Stoica, R. Morris, D. Karger, M.F. Kaashoek, and H. Balakrishnan, "Chord: A Scalable Peer-to-Peer lookup Service for Internet Applications," *ACM SIGCOMM*, 2001.

[34] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis*. Chapman and Hall, 1995.

[35] W. Willinger, M.S. Taqqu, R. Sherman, and D.V. Wilson, "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," *ACM SIGCOMM*, 1995.

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