UNDERSTANDING CHURN IN DECENTRALIZED PEER-TO-PEER NETWORKS

A Dissertation

by

ZHONGMEI YAO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2009

Major Subject: Computer Science

UNDERSTANDING CHURN IN DECENTRALIZED PEER-TO-PEER NETWORKS

A Dissertation

by

ZHONGMEI YAO

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Dmitri Loguinov
Committee Members,	Riccardo Bettati
	Jennifer L. Welch
	Narasimha Annapareddy
Head of Department,	Valerie E. Taylor

August 2009

Major Subject: Computer Science

ABSTRACT

Understanding Churn in
Decentralized Peer-to-Peer Networks. (August 2009)
Zhongmei Yao, B.S., Donghua University;
M.S., Louisiana Tech University
Chair of Advisory Committee: Dr. Dmitri Loguinov

This dissertation presents a novel modeling framework for understanding the dynamics of peer-to-peer (P2P) networks under churn (i.e., random user arrival/departure) and designing systems more resilient against node failure. The proposed models are applicable to general distributed systems under a variety of conditions on graph construction and user lifetimes.

The foundation of this work is a new churn model that describes user arrival and departure as a superposition of many periodic (renewal) processes. It not only allows general (non-exponential) user lifetime distributions, but also captures heterogeneous behavior of peers. We utilize this model to analyze link dynamics and the ability of the system to stay connected under churn. Our results offers exact computation of user-isolation and graph-partitioning probabilities for any monotone lifetime distribution, including heavy-tailed cases found in real systems. We also propose an age-proportional random-walk algorithm for creating links in unstructured P2P networks that achieves zero isolation probability as system size becomes infinite. We additionally obtain many insightful results on the transient distribution of in-degree, edge arrival process, system size, and lifetimes of live users as simple functions of the aggregate lifetime distribution.

The second half of this work studies churn in structured P2P networks that are usually built upon distributed hash tables (DHTs). Users in DHTs maintain two types

of neighbor sets: routing tables and successor/leaf sets. The former tables determine link lifetimes and routing performance of the system, while the latter are built for ensuring DHT consistency and connectivity. Our first result in this area proves that robustness of DHTs is mainly determined by zone size of selected neighbors, which leads us to propose a min-zone algorithm that significantly reduces link churn in DHTs. Our second result uses the Chen-Stein method to understand concurrent failures among strongly dependent successor sets of many DHTs and finds an optimal stabilization strategy for keeping Chord connected under churn.

To my family

ACKNOWLEDGMENTS

It is my great fortune that I have been working with my advisor and mentor Dmitri Loguinov at the Internet Research Lab (IRL) who has been the most important to my development as a computer scientist and the completion of this dissertation. Over these years, I have continuously been amazed by his invaluable guidance and incomparable wisdom. I owe immense thanks to him for fully supporting me and giving me the most exciting, enjoyable, and unforgettable experience at IRL. I would like to thank Daren B.H. Cline whose classes and weekly meetings have proven to be one of my best learning experiences at Texas A&M University (TAMU). Both of them have long been inspirations to me. Their excellence in research and teaching will guide me in my future academic career.

I am indebted to my committee members Riccardo Bettati, Jennifer L. Welch, and A. L. Narasimha Reddy for constantly supporting me through this thrilling and challenging journey. My gratitude also goes to Jason H. Li and Jianer Chen for their unwavering help and encouragements in my research career.

I would like to thank Xiaoming, Yueping, Derek, Hsin-Tsang, and Fenghui who are like my bothers and have provided help and friendship at times when it was most needed. I am thankful for the great times I have had with Clint, Seong, Chandan, Brad, Matt, and all other IRL members. I also owe gratitude to Dongxiao, Jie, Lili, Min, and Qiujie at TAMU for their previous help and friendship.

My heartfelt thanks go to my parents and parents in law for their love and for helping me take care of my daughter. I would not have achieved anything without their support. I am also deeply grateful to my brother and sister who always stand behind me.

I would like to thank my husband Weisong and daughter Sophie for giving me

the most wonderful love and the sweetest home. They are the miracles I have had. They are the sunshine in my life. They give me new dreams to pursue.

Finally, I extend my sincere thanks to anonymous reviewers of IEEE ICNP, IEEE INFOCOM, and IEEE/ACM Transactions on Networking and P. Brighten Godfrey for providing insightful comments on earlier versions of this work.

TABLE OF CONTENTS

CHAPTER Page

I INTRODUCTION 1

1.1. Research Problem 1
1.1.1 Background 3
1.2. Contributions 4
1.2.1 Modeling Foundation 4
1.2.2 Churn in Unstructured P2P Networks 6
1.2.3 Churn in Structured P2P Networks 7
1.3. Dissertation Structure 9

II RELATED WORK 11

2.1. Basics of P2P Graphs 11
2.2. Churn Models 13
2.3. Resilience in Unstructured P2P Networks 14
2.4. Link Dynamics in DHTs 15
2.5. Resilience of DHTs 16

III HETEROGENEOUS USER CHURN 17

3.1. Churn Model 17
3.1.1 Assumptions 18
3.1.2 Properties 19
3.1.3 Aggregate Lifetimes 22
3.2. Characteristics of Selected Users 26
3.2.1 Definitions 26
3.2.2 General Case 27
3.2.3 Uniform Selection 29
3.2.4 Lifetime of Users in the System 35
3.3. Summary 36

IV NODE OUT-DEGREE AND AGE-BASED NEIGHBOR SELECTION* 37

4.1. Introduction 37
4.1.1 Chapter Structure and Contributions 37
4.2. General Node Isolation Model 40

CHAPTER Page

4.2.1 Background 40
4.2.2 Hyper-Exponential Approximation 41
4.2.3 Isolation Probability 46
4.2.4 Verification of Isolation Model 50
4.2.5 Necessity of Neighbor Replacement 53
4.2.6 Discussion 56
4.3. Max-Age Selection 58
4.3.1 Residual Lifetime Distribution 58
4.3.2 Isolation and Resilience 65
4.4. Age-Proportional Neighbor Selection 67
4.4.1 Random Walks on Weighted Directed Graphs 67
4.4.2 Residual Lifetime Distribution 69
4.4.3 Isolation and Resilience 73
4.5. Summary 76

V NODE IN-DEGREE AND JOINT IN/OUT-DEGREE 78

5.1. Introduction 78
5.2. Edge Arrival 79
5.2.1 Definitions 80
5.2.2 Edge Creation Process 82
5.2.2.1 Uniform Integrability 83
5.2.2.2 Residuals 85
5.2.2.3 Edges 86
5.2.3 Edge Arrival Process 87
5.2.3.1 Continuity 88
5.2.3.2 Mean Convergence 88
5.2.3.3 Probability Convergence 89
5.2.4 Simulations 91
5.3. In-Degree 92
5.3.1 Expected In-Degree 92
5.4. Joint In/Out-Degree Model 96
5.4.1 Preliminaries 96
5.4.2 Exponential Lifetimes (Exact Model) 97
5.4.3 Isolation with Increased Age 100
5.4.4 Exponential Lifetimes (Asymptotic Model) 101
5.5. Summary 105

VI LINK LIFETIMES IN DHTS 106

CHAPTER Page

6.1. Introduction . . . 106
6.1.1 Analysis of Existing DHTs . . . 107
6.1.2 Improvements . . . 108
6.2. General DHT Model . . . 109
6.2.1 Assumptions . . . 110
6.2.2 Neighbor Dynamics . . . 110
6.3. Link Lifetime Model . . . 113
6.3.1 Preliminaries . . . 113
6.3.2 Neighbor Dynamics . . . 114
6.3.3 Conditional Link Lifetimes . . . 118
6.4. Deterministic DHTs . . . 121
6.4.1 Residual Lifetimes of Neighbors . . . 121
6.4.2 Exponential Lifetimes . . . 123
6.4.3 Pareto Lifetimes . . . 127
6.4.4 Zone Sizes . . . 130
6.4.5 Putting the Pieces Together . . . 133
6.5. Randomized DHTs . . . 136
6.5.1 Max-Age Selection . . . 136
6.5.2 Min-Zone Selection . . . 137
6.6. Summary . . . 141

VII SUCCESSOR LISTS IN DHTS . . . 142

7.1. Introduction . . . 142
7.1.1 Static Failure . . . 143
7.1.2 Dynamic Failure . . . 144
7.2. Static Node Failure . . . 145
7.2.1 Basic Asymptotic Model . . . 146
7.2.2 Discussion . . . 148
7.3. Dynamic Node Failure: General Results . . . 151
7.3.1 Successor List Model . . . 151
7.3.2 Node Isolation . . . 153
7.3.3 Closed-Form Bounds on ϕ . . . 155
7.3.4 Graph Disconnection . . . 159
7.4. Dynamic Node Failure: Effect of Stabilization Intervals . . . 163
7.4.1 Uniform Stabilization Delays . . . 164
7.4.2 Constant Stabilization Delays . . . 166
7.4.3 Optimal Strategy . . . 168
7.5. Heavy-tailed Lifetimes . . . 170

CHAPTER Page

7.6. Summary 170

VIII CONCLUSION AND FUTURE WORK 172

8.1. Conclusion 172
8.2. Future Work 174

REFERENCES 175

VITA 186

LIST OF TABLES

TABLE		Page
I	Comparison of model ϕ to simulations under uniform selection with $E[L] = 0.5$ hours and $k = 7$	52
II	Exact model (167) and simulations ($n = 2000$, $E[L] = 0.5$ hours)	100
III	Convergence of (178) to (167) for $E[L] = 0.5$ Hours and $k = 6$	105
IV	Comparison of simulation results of $P(X = 0)$ under static node failure to model (247) in Chord	149
V	Comparison of the asymptotic model (258) to the exact model (255) of node isolation probability ϕ with $E[L] = 0.5$ hours, $\rho = E[L]/E[S]$, and $r = 8$	159
VI	Comparison of model (279) 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 $E[S] = E[L]/\rho$.	163
VII	Convergence of simulation results to model $\phi_u/\phi = .0127$ from (281) for $E[L] = 0.5$ hours, $r = 6$, and $\rho = E[L]/E[S]$	166
VIII	Convergence of simulation results to model $\phi_c/\phi = .0014$ from (289) for $E[L] = 0.5$ hours, $r = 6$, and $\rho = E[L]/E[S]$	168

LIST OF FIGURES

FIGURE		Page
1	Example of a P2P network, where edges connecting peers are virtual links (dashed lines), e.g., pointers to IP addresses and port numbers of neighbors.	2
2	Structure of the dissertation.	9
3	User v's successors and neighbors in Chord.	12
4	Process $\{Z_i(t)\}$ depicting ON/OFF behavior of user i.	18
5	Sample path and distribution of $N(n,t)$ in system $\mathcal{H}$ with $n = 1000$ users. The Gaussian fit is from Lemma 1 after 10^6 iterations. .	22
6	Comparison of simulation results of $F(n,x)$ to model (9) in a graph with $n = 1000$ nodes. System evolved to age 10^5 hours.	25
7	Comparison of simulation results of $H(n,x)$ to model (39) in a graph with $n = 1000$ nodes. System age 500 hours and 10^5 iterations.	34
8	Impact of shape parameter α on model ϕ under uniform selection, Pareto lifetimes, $E[L] = 0.5$ hours, $\beta = (\alpha-1)E[L]$, exponential search delays, and $k = 7$.	54
9	Convergence of simulation results to model ϕ in (83) as system age $\mathcal{T} \to \infty$ under uniform selection, no neighbor replacement, and Pareto lifetimes with $\beta = (\alpha-1)E[L]$ in a graph with $n = 1,000$ nodes.	56
10	Accuracy of models (100) and (115) for Pareto lifetimes with $E[L] = 0.5$ hours and $\alpha = 3$ in a graph with $n = 5,000$ nodes.	63
11	Comparison of model ϕ to simulations using the max-age selection strategy for Pareto lifetimes with $E[L] = 0.5$ hours and $\alpha = 3$, exponential search times and $k = 7$ in a graph with $5,000$ nodes. . .	65

FIGURE Page

12 Influence of m on model ϕ under max-age selection for Pareto lifetimes with $E[L] = 0.5$ hours, exponential search times with $E[S] = 6$ minutes, and $k = 7$. 67

13 Comparison of model ϕ to simulations under age-proportional random walks for Pareto lifetimes, $E[L] = 0.5$ hours, $\beta = (\alpha-1)E[L]$, exponential search delays, and $k = 7$ in a graph with $n = 8,000$ nodes. 72

14 Impact of α on ϕ under uniform selection and under age-proportional random walks for Pareto lifetimes, $E[L] = 0.5$ hours, $\beta = (\alpha - 1)E[L]$, exponential search delays, and $k = 7$. 74

15 Simulation results of ϕ under age-proportional selection as system age $\mathcal{T}$ and size n increase for Pareto lifetimes with $E[L] = 0.5$ hours. 76

16 Process $\{Y_i^c(t)\}$ indicates DEAD/ALIVE behavior of the c-th outlink of user i. Process $\{U_i^c(t)\}$ counts the number of DEAD→ALIVE transitions within the current ON cycle of i. 81

17 Distribution of edge inter-arrival delays approaches exponential with rate ν in (147) for $k = 10$ and $\theta = 10$ using 10^9 iterations. . . . 92

18 Distribution of the number of edge arrivals to a node in the interval $[t, t + \Delta t]$ in a system with $n = 1000$ users, $k = 10$, and $\theta = 10$. The lines show Poisson fits with ν in (147) at $t = 500$ hours and after 10^5 iterations. 93

19 Comparison of the model for $E[X(t)]$ to simulation results for $n = 2000$, $E[L] = 0.5$ hours, and $k = 8$ after 10^6 iterations. 95

20 The CDF of T_{out} and T for exponential lifetimes with $E[L] = 0.5$ hours, exponential search delays with $E[S] = 0.1$ hours, and $k = 6$. 101

21 User v's neighbors in the DHT. 111

22 The i-th link failure and replacement of user v who joins at time 0 in a DHT, $1 \leq i \leq k$. 112

23 Zone size U and remaining zone size Y_j of user u. 114

FIGURE Page

24 State diagram for the process $\{A_\delta^j, \delta \geq 0\}$ of neighbor changes. 115

25 Comparison of simulation results to model (205) in a deterministic DHT with $E[N] = 1,000$. In both cases, $E[L] = 1$ hour. 123

26 Comparison of model (207) to simulations in a deterministic DHT with $E[N] = 2,000$ and exponential user lifetimes with $E[L] = 1$ hour. 126

27 Comparison of model $E[R(y)]$ in Theorem 18 to simulation results in a deterministic DHT with mean size $E[N] = 2,000$ and Pareto user lifetimes L with mean $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$. . . . 129

28 Comparison of simulation results of Y_j to model (231) in a deterministic DHT with mean size $E[N] = 500$ under churn produced by Pareto L with $\alpha = 3$ and $E[L] = 1$ hour. 133

29 Comparison of $E[R_j]$ to $E[Z_j]$ in a deterministic DHT with mean size $E[N] = 2,500$ users, Pareto lifetimes with mean $E[L] = 1$ hour, and $\beta = E[L](\alpha - 1)$. 134

30 Link lifetimes R_4 are less heavy-tailed than Pareto user lifetimes L in a deterministic DHT with mean size $E[N] = 2,500$ peers, $E[L] = 1$ hour, and $\beta = (\alpha - 1)E[L]$. 135

31 Impact of shape α and number of samples m on mean link lifetime $E[R_j]$ under max-age selection in a randomized DHT with mean size $E[N] = 2,000$ for Pareto lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$. 137

32 Comparison of mean link lifetime $E[R_j]$ under min-zone selection to that under max-age selection in a randomized DHT with mean size $E[N] = 2,000$ for Pareto user lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$. 138

33 Approximation of $E[R_j]$ as a linear function of number of samples m under min-zone selection for Pareto user lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$. 140

34 Evolution of a node's successor list over time. 152

35 Markov chain $\{Z(t)\}$ modeling a node's successor list. 154

FIGURE Page

36 Comparison of model (255) to simulation results on node isolation probability ϕ for exponential lifetimes with $E[L] = 0.5$ hours and exponential stabilization intervals with $E[S] = E[L]/\rho$. 156

37 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. 171

CHAPTER I

INTRODUCTION

1.1. Research Problem

Peer-to-peer (P2P) networks are a recently emerged distributed architecture, in which all participants (i.e., peers) in the network often supply resources (e.g., bandwidth, storage, and computing power) to each other and simultaneously serve as both servers and clients. The most salient characteristic of these systems is that communication between users takes place directly instead of relying on central servers (see Fig. 1 for an example). By utilizing resources at the edge of the Internet, P2P networks have become an efficient and scalable platform for distributed applications (e.g., file sharing, media streaming, and telephony) that support millions of users online. More significantly, the power of P2P computing may soon revolutionize our computing experience and reinvent the essence of data transfer in the Internet over the next ten years [86].

Unlike other distributed systems where failures may be considered rare or abnormal, most P2P networks constantly remain in the state of *churn*, which is a general term describing dynamic behavior of these systems in which arrival/failure of individual users are not synchronized. The analysis of how these systems behave during churn has recently attracted significant attention and has become an important research area [6], [16], [29], [26], [33], [34], [39], [40], [43], [46], [50], [61], [66].

While many properties of a system (e.g., throughput, load-balancing, efficiency

---

The journal model is *IEEE/ACM Transactions on Networking*.

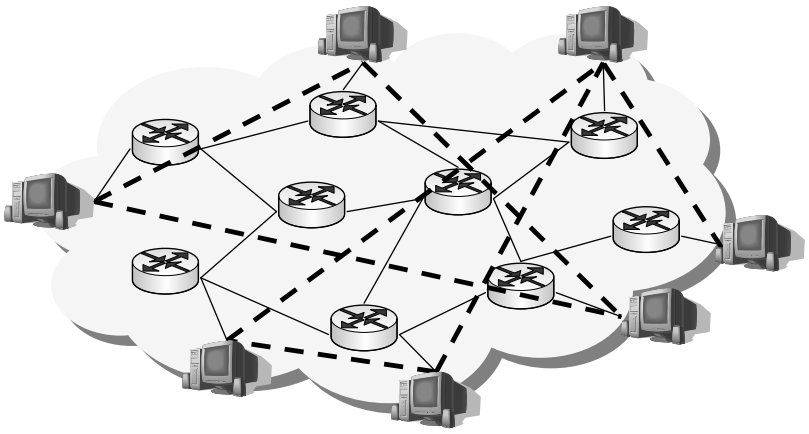


Fig. 1. Example of a P2P network, where edges connecting peers are virtual links (dashed lines), e.g., pointers to IP addresses and port numbers of neighbors.

of routing, message overhead, and file popularity) affect its usefulness to the user, we focus in this work on *resilience* of P2P networks, which is defined as the ability of the network to continuously provide services when the system experiences churn. In *centralized* P2P systems (e.g, Napster [75] and a swarm with a centralized tracker in BitTorrent [66]) where central servers have a global view of the system and respond to certain types of requests (e.g. content-search), the issue of resilience reduces to the single point of failure problem. In contrast, *decentralized* P2P networks (e.g, Gnutella [25], KaZaA [35], Chord [79], and Kademlia [58]) have no single point of failure and embrace frequent node failures as part of their normal operation. The goal of this work is to offer generic models for *understanding churn in these decentralized P2P networks and designing systems more resilient against node failure.*

Recall that (decentralized) P2P networks organized users into distributed graphs that provide system-wide services by routing requests between neighboring nodes. As a result, two fundamental issues in these decentralized networks are understanding link dynamics (i.e., delay between formation and failure of each link) and ability of the system to stay connected under churn [3], [6], [26], [29], [34], [39], [40], [41], [43],

[44], [50], [61], [79]. However, before resilience and performance of P2P networks can be fully understood, a good model of churn is required since even today most analytical models that consider churn [39], [43], [50], [61] do not capture the inherent heterogeneity of users or the behavior of P2P networks under non-exponential lifetimes.

1.1.1 Background

In many P2P networks, each user v creates k links to other peers when joining the system, where k may be a constant or a function of system size [52], and detects/repairs failed links in order to remain connected and perform P2P tasks (e.g., routing and key lookups) [67], [71], [72], [79]. This type of churn was originally formalized in [43], where Leonard *et al.* equipped joining users with random lifetimes L_i that determined the duration of their presence in the system and modeled neighbor replacement using random delays S_i that included the timeouts to detect each neighbor failure and protocol delays to actually obtain a new neighbor. Given this setup, link behavior is often modeled as an ON/OFF process in which each link is either ON at time t, which means that the corresponding user is currently alive, or OFF, which means that the user adjacent to the link has departed from the system and its failure is in the process of being detected and repaired. ON durations of links are commonly called *link lifetimes* R_i and their OFF durations are *repair delays* S_i that included the timeouts to detect each failure and protocol delays to actually obtain a new neighbor. The out-degree of a live user is simply the number of links that are in the ON state.

With this setup, it is not hard to see that characterizing link dynamics is fundamental to understanding the behavior of P2P networks since it directly affects resilience, performance, and reliability of P2P networks. For instance, longer average link lifetime means that users must repair failed links less frequently, which leads to

smaller churn rates in the terminology of [26], and queries are less likely to encounter dead neighbors during routing [39], which yields larger data delivery ratios [84] and higher lookup success rates. This model [43], however, treated P2P users equally in their online characteristics (i.e., all user lifetimes were drawn from the same distribution), did not capture the impact of in-degree on the resilience of the system, and did not consider different neighbor replacement phenomena in unstructured and structured P2P implementations.

1.2. Contributions

The foundation of this dissertation is a new user churn model in P2P systems. We later utilize this model to understand the dynamics in both unstructured and structured P2P networks under a variety of conditions on user lifetimes and neighbor selection strategies.

1.2.1 Modeling Foundation

Heterogeneity of lifetimes is a fundamental property of P2P systems where some users consistently spend substantial periods of time in the system and others very little [81]. This observation prompts the question of *whether P2P systems can indeed be modeled using a single homogeneous lifetime distribution without sacrificing model accuracy*? In addition to lifetimes, churn is characterized by the distribution of offline durations, which together with lifetimes define the *availability* of each user [8], [74], i.e., the average fraction of time a user is logged in. It is therefore important to understand how offtimes contribute to the dynamics of the system and which peer characteristics affect local graph-theoretic properties (e.g., distribution of in and out-degree at each time t, probability that a given neighbor is alive, isolation probability

within a lifetime) of each user.

To answer these questions, we offer a generic churn model that captures the *heterogeneous* behavior of end-users, including their difference in online habits and diversity of offline "sleep time." We view each user as an alternating renewal process that is ON when the user is logged in and OFF otherwise, where online/offline durations of each user i are respectively drawn from distributions $F_i(x)$ and $G_i(x)$. This approach creates a system of *heterogeneous* users, each with its own profile of behavior that stays constant during the peer's recurring participation in the network [81].

Armed with this model, we obtain the aggregate lifetime distribution $F(x)$ of all users who have joined the system, the lifetime distribution $J(x)$ of the users currently online, and the residual lifetime distribution $H(x)$ of a randomly selected user in the network. Our results show that all three metrics are weighted functions of individual lifetime distributions $F_i(x)$, where $H(x)$ is additionally dependent on the number of users currently in the network, the probability that a given user is picked by joining peers, and the conditional residual lifetimes of neighbors chosen by the selection method. The model for $H(x)$ is extremely complex and generally intractable unless neighbor selection is performed *uniformly among currently participating users* (e.g., by picking users from uniformly random subsets of cached nodes or using special random walks on the graph [99]), in which case we show that $H(x)$ can be directly obtained from $F(x)$. This is an important conclusion that demonstrates that instead of measuring n individual lifetime distributions, where n is the total number of users participating in the system, one can measure lifetimes of joining users to obtain $F(x)$, which is then *sufficient to entirely model the effect of churn* on unstructured P2P graphs.

We also revisit the observation of [81] that the users already present in Gnutella

and BitTorrent networks exhibit larger average lifetimes than those joining the system. We show that this effect is a consequence of $J(x)$ being the *spread* [91] of distribution $F(x)$, which allows us to prove that random users currently in the system have stochastically larger lifetimes than random arriving users *regardless of the shape of distributions* $F_i(x)$ and $G_i(x)$. We additionally show that while $F(x)$ may appear to be heavy-tailed as observed in practice [12], [30], [47], it is possible that individual lifetime distributions $F_i(x)$ may *all* be exponential, or contain a mix of exponential and heavy-tailed distributions. Occurrence of this effect depends on random availability of each user and shows that conclusions on the individual habits of peers may not be drawn from their aggregate behavior $F(x)$.

1.2.2 Churn in Unstructured P2P Networks

Users in unstructured P2P systems (e.g., Gnutella [25], KaZaA [35]) rely solely on their routing tables (i.e., sets of link pointers) to provide system-wide services to each other. One of the primary metrics of resilience is *graph disconnection* during which a P2P network partitions into several non-trivial subgraphs and starts to offer limited service to its users. However, as shown in our early work [44], most partitioning events in well-connected P2P networks are single-node isolations, which occur when all neighbors in the routing table of a node v are in the failed status before v is able to detect their departure and then replace them with other alive users. For such networks, node isolation analysis has become the primary method for quantifying network resilience in the presence of user churn.

Traditional analysis of node isolation and graph partitioning in unstructured P2P networks [42], [61] have assumed exponential user lifetimes and only considered age-independent neighbor replacement. In this dissertation, we overcome these limitations by introducing a general node-isolation model for heavy-tailed user lifetimes and

arbitrary neighbor-selection algorithms. Using this model, we analyze two age-biased neighbor-selection strategies and show that they significantly improve the residual lifetimes of chosen users, which dramatically reduces the probability of user isolation and graph partitioning compared to uniform selection of neighbors. In fact, the second strategy based on random walks on age-proportional graphs demonstrates that for lifetimes with infinite variance, the system monotonically *increases* its resilience as its age and size grow. Specifically, we show that the probability of isolation converges to zero as these two metrics tend to infinity. We conclude the part with simulations in finite-size graphs that demonstrate the effect of this result in practice.

The above approach only models the *out-degree* of each user and does not consider the increased resilience arising from additional *in-degree* edges arriving in the background to each user during its stay in the system. We overcome this shortcoming and build a complete closed-form model characterizing the evolution of in-degree in unstructured systems under the assumption of uniform neighbor selection. We formally prove that despite node heterogeneity and non-Poisson arrival dynamics, the edge-arrival process to each user approaches Poisson as system size becomes sufficiently large. This allows relatively simple analytical treatment of the edge-arrival process and leads to closed-form results on the transient distribution of in-degree as a function of the general user lifetime distribution. We finish the part by combining the in and out-degree isolation models into a single approximation that clearly shows the contribution of in-degree to the resilience of the graph.

1.2.3 Churn in Structured P2P Networks

Unlike unstructured P2P graphs where nodes have more autonomy to choose neighbors, structured P2P networks that are usually built upon Distributed Hash Tables (DHTs) have limited choices to build edges. DHTs (e.g., Chord [79], Kademlia [58],

CAN [67], and Pastry [72]) provide a lookup service similar to hash tables, but the task of storing (key, value) pairs is distributed among users in the system. Nodes in DHTs maintain *routing tables* and *successor/leaf sets* to ensure that any peer can efficiently route a search request to the node that is responsible for the desired key. In particular, routing tables determine link lifetimes and general routing performance of the system, while successor sets are built for ensuring DHT consistency (so that the system guarantees that all lookups are resolved correctly) and keeping the system connected. While it was known that P2P system performance depended mainly on link lifetimes and that successor lists were essential to DHT consistency, there were no frameworks or even high-level approaches for studying these neighbor sets in DHTs under churn.

In DHTs, link lifetimes are rather complicated since links actively switch to new neighbors before current neighbors die in order to balance the load and ensure DHT consistency. To understand neighbor churn in such networks, we propose a simple, yet accurate, model for capturing link dynamics in structured P2P systems and obtain the distribution of link lifetimes for fairly generic DHTs. Similar to [26], our results show that deterministic networks (e.g., Chord [79], CAN [67]) unfortunately do not extract much benefit from heavy-tailed user lifetimes since link durations are dominated by small remaining lifetimes of newly arriving users that replace the more reliable existing neighbors. We also examine link lifetimes in randomized DHTs equipped with multiple choices for each link and show that users in such systems should prefer neighbors with *smaller zones* rather than larger age as suggested in prior work [45], [84]. We finish this analysis by demonstrating the effectiveness of the proposed min-zone neighbor selection for heavy-tailed user lifetime distributions with the shape parameter α obtained from recent measurements [12], [89].

The second neighbor set of each user in DHTs is the successor list consisting

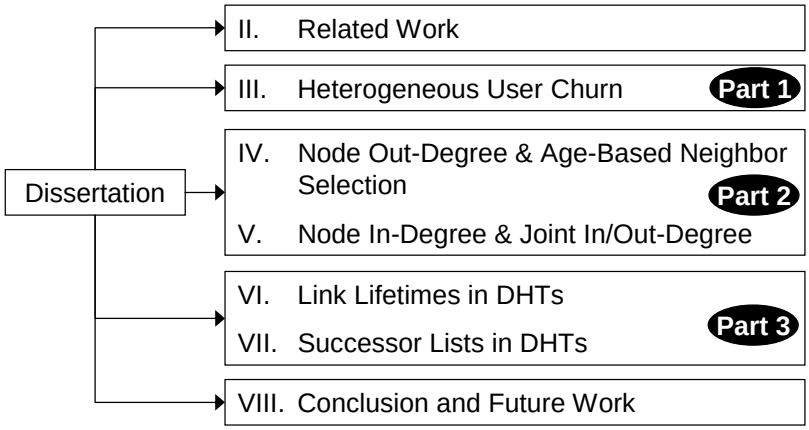


Fig. 2. Structure of the dissertation.

of peers that immediately follow it in the DHT key space. Periodic stabilizations keep the successor list up to date. Successor lists are essential to structured P2P networks because the system becomes disconnected as soon as the entire successor list of *any* node fails. The main difficulty in analyzing this disconnection problem is that successor lists of consecutive users in the DHT key space exhibit strong *dependency*. We apply the Erdös and Rényi law and the Chen-Stein method to derive closed-form results on the probability of partitioning in Chord under both static and dynamic node failure and find an optimal stabilization strategy for keeping Chord connected when the system experienced churn.

1.3. Dissertation Structure

The structure of the rest of this dissertation is shown in the following.

As illustrated in Fig. 2, Chapter II overviews P2P networks and the state of the art of the analytical work on the dynamics of these networks. Chapter III introduces our modeling foundation, i.e., heterogeneous user churn model, and studies three important distributions that are later used for analyzing churn in unstructured and

structured P2P networks. The second part of this work focuses on churn in unstructured P2P networks. Chapter IV presents a new generic node isolation model under various neighbor selection strategies and for non-exponential user lifetimes. We further propose an age-proportional random-walk algorithm for selecting neighbors. In Chapter V, we derive closed-form results on the transient distribution of in-degree as a function of the user lifetime distribution and then examine the joint in/out-degree model.

The third part of this dissertation studies churn in DHTs. Chapter VI analyzes link dynamics in classic DHTs and finds that zone sizes play a key role in determining link lifetimes. This leads us to the min-zone selection algorithm which significantly improve the robustness of DHTs. In Chapter VII, we study successor lists in DHTs that are used to ensure graph connectivity and DHT consistency. We conclude this dissertation and discuss the future work in Chapter VIII.

CHAPTER II

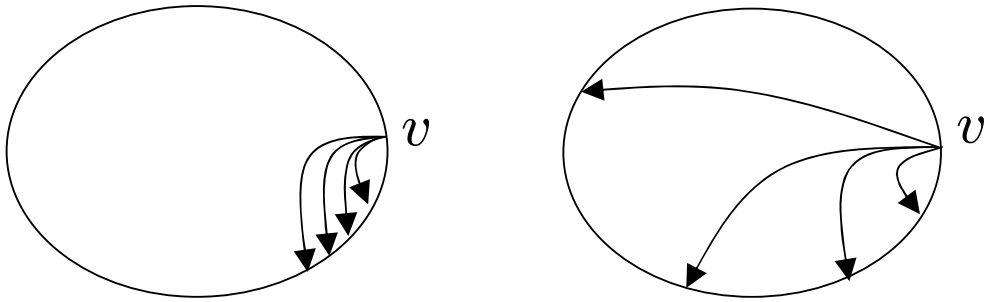
RELATED WORK

2.1. Basics of P2P Graphs

P2P networks can be broadly classified as unstructured and structured [54]. As their names imply, the former systems organize users onto random graphs, while the latter graphs are constructed based on fixed rules, where nodes' links share common structured patterns.

Many popular unstructured P2P networks, including Gnutella [25] [82], KaZaA [35] [49], BitTorrent [9], [31], [65], [66] support keyword-based searches. In these systems, nodes usually use flooding, random walks [23], or hybrid methods [24] to route requests until some users that have the desired content are reached. Search is often efficient only for popular content. To improve routing efficiency, Gnutella, KaZaA, and Skype [28], [76] utilize the supernode and peer hierarchical structure (i.e., parent-children structure) and organize supernodes onto decentralized graphs. Supernodes resolve/forward queries for their children. In BitTorrent, a centralized tracker is used to find peers that have the desired file. Other approaches without relying on centralized servers will be discussed in Section 2.3.

Existing structured P2P networks that are developed on DHTs support efficient exact key lookups [13], [29], [34], [55], [56], [67], [71], [80], [92]. They map keys of data items and peers into the same identifier (ID) space (e.g., continuous space $[0, 1)$ or discrete set $\{0, 1, \ldots, 2^{64} - 1\}$) and assign each content's key to a set of peers whose IDs are closest to that key. Unlike unstructured graphs, DHTs have the coupling between keys of data items and peers on the graph and thus ensure that queries are



(a) r successors (b) k finger table links

Fig. 3. User v's successors and neighbors in Chord.

resolved. We use Chord as an example to understand the basics of DHTs.

Chord [80] maps each node and key using a uniform hashing function into the identifier (ID) space $\{0, 1, \ldots, 2^m - 1\}$, where m is some sufficiently large number that can accommodate all nodes without conflict. 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. 3, each user v in Chord builds a *successor list* and a *finger table*. Assuming n users in the system, the former set contains $r = \Theta(\log n)$ peers immediately following user v along the ring and the latter set consists of $k = \Theta(\log n)$ neighbor pointers where the i-th neighbor is the owner of the key $id(v) + 2^i$.

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, 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 sustain 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 alive peer in its successor list, the system is able to correct (after some delay) all stale finger pointers and re-populate 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, that user is considered *isolated* and Chord becomes *partitioned* [80]. Recovery from such disconnection is not guaranteed in the general case.

2.2. Churn Models

One of the first models of churn was proposed in [61], which assumed an unstructured P2P system with Poisson arrivals and departures that could be modeled as an $M/M/1$ queue. Neighbor replacement in this system was in direct response to failures and was assumed to be instantaneous, where the possibilities for replacement were limited to the nodes currently alive in a certain centralized cache. The paper showed that under user churn the graph remained connected and exhibited a logarithmic diameter, both with high probability.

Later models of churn [50] and recently [39] assumed a DHT-like system in which repair algorithms were run *independently* of user failures and at exponentially distributed intervals (i.e., as Poisson processes). This approach modeled the consis-

tency check algorithm in Chord, which periodically verified the successor list and corrected invalid pointers. These models assumed homogeneous exponential lifetimes and Poisson arrival/departure processes with no way of generalizing their results to non-exponential system dynamics.

A different approach was undertaken in [43], where neighbor replacements were explicitly initiated in response to failed links. In this setup, each joining user randomly selected k neighbors from the graph and then monitored their online presence using keep-alive messages. Once the failure of an existing neighbor was detected, a uniformly random replacement was sought from among the currently alive users in the system. Detection and replacement delays were also random, but explicitly non-zero. Under these conditions, the paper showed that each user became isolated with probability no larger than $\phi_{out} = k\rho/(1+\rho)^k$, where ρ was the ratio of the average lifetime to the average replacement delay, for all lifetime distributions with an exponential or heavier tail. This result was later generalized in [44] to show that the probability of non-partitioning in many P2P networks converged as $n \to \infty$ to that of avoiding isolation for each online user.

2.3. Resilience in Unstructured P2P Networks

Construction and maintenance of overlay networks consists of initial neighbor selection and subsequent replacement of dead links. Many P2P systems, including structured [13], [34], [47], [58], [63], [67], [72], [79], [98], and unstructured [15], [57], [61], [73], [88], perform neighbor selection and replacement to achieve the desired routing efficiency and search performance in the face of node joins and departures.

Gnutella, for example, sends a ping message every 3 seconds and detects link failure when TCP declares the connection aborted, which happens after several (e.g., 5 in Windows) subsequently failed retransmission attempts.

Previous work has used proximity-based neighbor selection to reduce lookup latency [29], [57], [68], [97], capacity-based selection to improve system scalability [15], [41], [78], and age-biased neighbor preference to improve reliability of the system [12], [41], [58], [77]. Additional studies have analyzed the tradeoffs between resilience and proximity [16] as well as studied how well different neighbor selection and recovery strategies could handle churn in DHTs [26], [71]. In recent work [87], [88], random walks have been used to build unstructured P2P systems and replace failed links with new ones. Finally, only a handful of modeling studies of user isolation and neighbor selection under churn exist [39], [42], [50], [61]. They are mostly limited to exponential user lifetimes and age-unrelated user replacement and do not capture the effect of in-degree on resilience.

2.4. Link Dynamics in DHTs

Among the recent studies of link lifetimes, one direction focuses on non-switching P2P systems. Leonard *et al.* [42] show that heavy-tailed lifetimes allow link lifetime $E[R]$ to be significantly larger than user lifetime $E[L]$. Additional results of this model and its application to unstructured networks are available in [45], [93], [96]. Another recent study [84] examines DHTs without switching with a focus on the *delivery ratio*, which is the fraction of time that all forwarding nodes between each source and destination are alive. Their results show that the delivery ratio is a function of link lifetime R for all examined neighbor-selection techniques.

The other direction also covers switching networks exemplified by traditional DHTs. Godfrey *et al.* [26] study the impact of node-selection techniques on the churn rate and observe that switching DHTs exhibit dramatically smaller link lifetimes than non-switching networks. Krishnamurthy *et al.* [39] compute the probability that

neighbors in Chord are in one of three states (alive, failed, or incorrect) and use this model to predict lookup consistency and query latency.

Additional work [8], [13], [46], [47], [48], [71], [81] focuses on measurement and simulation of structured P2P systems under churn.

2.5. Resilience of DHTs

Performance of DHTs under p-fraction node failure [29], [34], [80] and churn [13], [39], [46], [48], [50], [58], [71] have received significant attention since the advent of structured P2P networks. While the problem of connectivity under failure for general graphs remains NP-complete [22], [36], [83], recent work [45] 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 [45] 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 [45].

Another modeling work by Krishnamurthy *et al.* [39] 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.

CHAPTER III

HETEROGENEOUS USER CHURN

3.1. Churn Model

To understand the dynamics of churn and performance of P2P systems, we start by creating a model of user behavior and specifying assumptions on peer arrival, departure, and selection of neighbors. The focus of this section is to formalize recurring user participation in P2P systems in a simple model that takes into account heterogeneous browsing habits and explains the relationship between the various lifetime distributions observable in P2P networks.

Consider a P2P system with n participating users, where each user i is either alive (i.e., present in the system) at time $t \geq 0$ or dead (i.e., logged off). This behavior can be modeled by an ON/OFF right-continuous process $\{Z_i(t)\}$ for each i:

$$Z_i(t) := \begin{cases} 1 & \text{user } i \text{ is alive at time } t \\ 0 & \text{otherwise} \end{cases}, \quad 1 \leq i \leq n. \tag{1}$$

This framework is illustrated in Fig. 4, where parameter m stands for the cycle number and random variables $L_{i,m} > 0, D_{i,m} > 0$ are durations of user i's ON (life) and OFF (death) periods, respectively. The figure also shows the *residual process* $R_i(t)$, which is the duration of user i's remaining online presence from time t conditioned on the fact that it was alive at t.

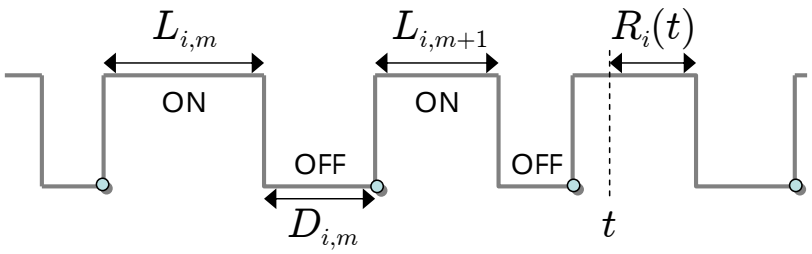


Fig. 4. Process $\{Z_i(t)\}$ depicting ON/OFF behavior of user i.

3.1.1 Assumptions

We next make several modeling assumptions about this system and explain how users generate their online/offline durations.

Assumption 1. *Set $\{Z_i(t)\}_{i=1}^n$ consists of mutually independent, alternating renewal processes.*

To elaborate, we restrict ON durations $\{L_{i,m}\}_{m=1}^\infty$ of user i to independent random variables (r.v.) with a general cumulative distribution function (CDF) $F_i(x)$ and OFF durations $\{D_{i,m}\}_{m=1}^\infty$ to independent r.v. with another CDF $G_i(x)$. This assumption also implies that the two sequences $\{L_{i,m}\}_{m=1}^\infty$ and $\{D_{i,m}\}_{m=1}^\infty$ are independent. We leave discussion of the more general case of correlated ON/OFF cycles to future work. Mutual independence in Assumption 1 additionally states that users do not synchronize their arrival or departures and generally exhibit uncorrelated lifetime characteristics (e.g., users simultaneously present in the system with multiple identities are not very common and have no large-scale impact on the dynamics of the network).

While Assumption 1 is a good start and allows certain results below to hold, asymptotically large systems require additional constraints on how users select their distributions $F_i(x), G_i(x)$. We next suppose that there are $\mathcal{T} \geq 1$ user types in the system representing different behavior (e.g., desktop peers that stay in the system for

days is one type, while laptop users that frequently disconnect is another). Before the network starts to evolve, each user randomly decides on its type, which then remains fixed for all $t > 0$.

Assumption 2. *(a) There exists some set $\mathcal{F}$ of distinct pairs of non-lattice CDFs defining non-negative random variables:*

$$\mathcal{F} := \left\{ \left(F^{(1)}(x), G^{(1)}(x)\right), \ldots, \left(F^{(\mathcal{T})}(x), G^{(\mathcal{T})}(x)\right) \right\},$$

where $\mathcal{T} \geq 1$ is a fixed number of user types. Further, each mean $l^{(j)} := \int_0^\infty (1 - F^{(j)}(x))dx$ and $d^{(j)} := \int_0^\infty (1 - G^{(j)}(x))dx$ satisfies $0 < l^{(j)}, d^{(j)} < \infty$ for all types $j = 1, \ldots, \mathcal{T}$;

(b) The pair of ON/OFF duration CDFs $(F_i(x), G_i(x))$ of each user i, $i = 1, \ldots, n$, is independently drawn from set $\mathcal{F}$, where type j is selected with probability (w.p.) $p_j \geq 0$ and $\sum_{j=1}^{\mathcal{T}} p_j = 1$;

(c) Defining $\mathcal{S}$ to be set of selections made by each user and conditioning on $\mathcal{S}$, Assumption 1 holds.

Assumption 2(a) uses $\mathcal{T}$ as the "diversity" factor of user behavior (e.g., $\mathcal{T} = 1$ reduces the system to a network of homogeneous users) and mandates that all average online/offline durations are both positive and finite. Part (b) allows for bias in the selection process and lets certain user types be more popular than others. Part (c) ensures that the system complies with Assumption 1 during its evolution. Note that Assumption 1 is more general and includes Assumption 2 as a special case.

3.1.2 Properties

We next explain the ON/OFF distributions commonly considered in this chapter and obtain basic properties of the system. The first lifetime distribution is exponential

$F_i(x) = 1 - e^{-\mu_i x}, \mu_i > 0$, with mean $1/\mu_i$. The second one is shifted Pareto

$$F_i(x) = 1 - (1 + x/\beta_i)^{-\alpha_i}, \quad \alpha_i > 1, \beta_i > 0, \tag{2}$$

with mean $\beta_i/(\alpha_i - 1)$. Offline distributions $G_i(x)$ do not affect our analysis and are kept general. For convenience of notation, define the mean lifetime of each user $l_i := E[L_{i,m}]$ and the mean offline duration $d_i := E[D_{i,m}]$, where the average is taken over all cycles $m = 1, 2, \ldots$ Denote the reciprocal of the mean ON/OFF cycle length of user i by

$$\lambda_i := (l_i + d_i)^{-1}, \tag{3}$$

which is the time-averaged arrival rate of the user into the system. We easily obtain from Smith's theorem that the asymptotic *availability* of each user i, i.e., the probability that it is in the system at an arbitrary instance t, is given by

$$a_i := \lim_{t\to\infty} P(Z_i(t) = 1) = \frac{l_i}{l_i + d_i}. \tag{4}$$

The final metric related to our churn model is the distribution of the number of users in the system. Denote by $N(n,t) := \sum_{i=1}^{n} Z_i(t)$ the number of users in the network at time t and notice that it is also a random process that fluctuates with time. Since many P2P properties of interest require stationarity, our analysis below is frequently confined to limiting distributions when network age $t \to \infty$, which we call *equilibrium*.

Define Z_i to be a Bernoulli r.v. with the equilibrium distribution of $Z_i(t)$, i.e., $P(Z_i = 1) = a_i$, where a_i is given in (4). Further define $N(n) := \sum_{i=1}^{n} Z_i$, which is a r.v. with the equilibrium distribution of $N(n,t)$. Based on Lyapunov's central limit theorem, it is easy to show that the equilibrium system size is approximately Gaussian for large n.

Lemma 1. *Under Assumption 2, we have as* $n \to \infty$

$$\frac{N(n) - \bar{N}_n}{\sigma_n} \xrightarrow{D} \mathcal{N}(0,1), \tag{5}$$

where $\bar{N}_n := \sum_{i=1}^{n} a_i$, $\sigma_n^2 := \sum_{i=1}^{n} a_i(1-a_i)$, *and* $\mathcal{N}(0,1)$ *denotes a standard normal r.v.*

Proof. The mean number of users alive in the equilibrium is

$$E[N(n)] = \sum_{i=1}^{n} E[Z_i] = \sum_{i=1}^{n} a_i, \tag{6}$$

which is the sum of all users' availability. Due to the independence among users, the variance of $N(n)$ is:

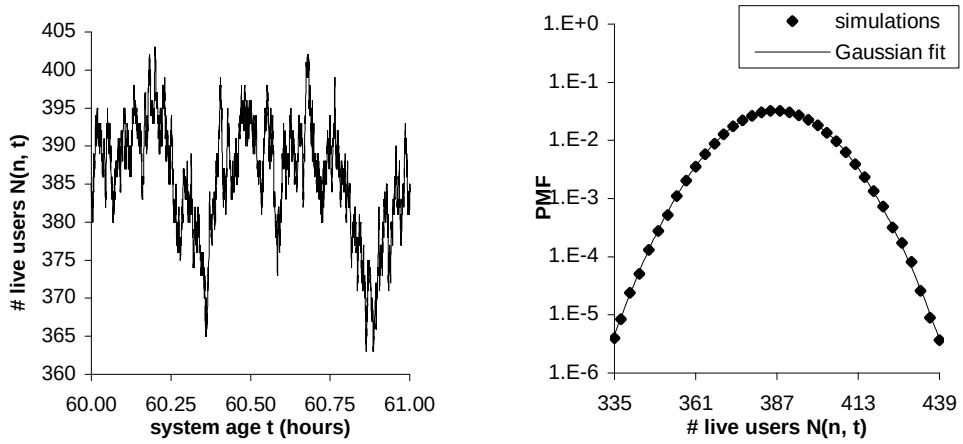
$$Var[N(n)] = \sum_{i=1}^{n} Var[Z_i] = \sum_{i=1}^{n} a_i(1-a_i). \tag{7}$$

Next, denote by m_{i2} the second central moment, and by m_{i3} the third central moment of Bernoulli variable $Z_i = \lim_{t\to\infty} Z_i(t)$. Since a_i are constants, it is easy to see that m_{i2} and m_{i3} are constants too. It immediately follows that

$$\frac{\left(\sum_{i=1}^{n} m_{i3}\right)^{1/3}}{\left(\sum_{i=1}^{n} m_{i2}\right)^{1/2}} \to 0, \tag{8}$$

showing that the Lyapunov condition of the Central Limit Theorem [62] holds. Thus, we conclude that the shifted and scaled $N(n)$ tends to a Gaussian r.v. as $n \to \infty$. □

We next show simulations explaining this result and its accuracy in systems with *finite* age and size. We generate a network of n users whose arrival/departure follows the introduced churn model. The system evolves for at least 50 virtual hours before being examined, which models non-trivial age of existing networks. We start by generating $\mathcal{T} = 1,000$ pairs of means $l^{(j)}$ and $d^{(j)}$, which are drawn randomly from



(a) evolution (b) PMF

Fig. 5. Sample path and distribution of $N(n,t)$ in system $\mathcal{H}$ with $n = 1000$ users. The Gaussian fit is from Lemma 1 after 10^6 iterations.

two Pareto distributions with $\alpha = 3$ as described next. For mean ON durations, we use $\beta = 1$ and obtain $E[l^{(j)}] = 1/2$ hour; for mean OFF durations, we use $\beta = 2$ and get $E[d^{(j)}] = 1$ hour. We study three cases throughout the chapter: 1) heavy-tailed system $\mathcal{H}$ with $F^{(j)}(x) \sim \text{Pareto}(3, 2l^{(j)})$ and $G^{(j)}(x) \sim \text{Pareto}(3, 2d^{(j)})$; 2) very heavy-tailed system $\mathcal{VH}$ with $F^{(j)}(x) \sim \text{Pareto}(1.5, l^{(j)}/2)$ and $G^{(j)}(x) \sim \text{Pareto}(1.5, d^{(j)}/2)$; and 3) exponential system $\mathcal{E}$ with $F^{(j)}(x) \sim \exp(1/l^{(j)})$ and $G^{(j)}(x) \sim \text{Pareto}(3, 2d^{(j)})$, where notation $\text{Pareto}(\alpha_i, \beta_i)$ refers to (2). The actual pairs $(F_i(x), G_i(x))$ are selected uniformly randomly from $\mathcal{F}$.

Fig. 5(a) shows one example for the evolution of system size $N(n,t)$ as a function time t. Part (b) of the figure shows the PMF (probability mass function) of $N(n,t)$ at $t \gg 0$ and a Gaussian fit from Lemma 1, confirming its accuracy.

3.1.3 Aggregate Lifetimes

Prior measurement studies [81], [89] sampled lifetimes of all joining users over some long period of time to characterize the dynamics of P2P systems. We are now inter-

ested in what metric they estimated and how it can be expressed in our notation. For each instance of user i being present in the system during interval $[0, t]$, place its ON duration $L_{i,m}$ into set $S_i(t)$ and define $S(t) = \cup_{i=1}^{n} S_i(t)$. Then let $F(n, t, x)$ be the CDF of values collected in set $S(t)$ (i.e., the probability that the obtained lifetimes are less than or equal to x). Finally, define $F(n, x) := \lim_{t\to\infty} F(n, t, x)$ to be the *aggregate lifetime distribution* of the system and $l(n)$ to be its mean (both exist from Assumption 2).

Our next result shows that $F(n, x)$ a weighted average of individual lifetime distributions, where the weights are biased toward those peers who frequently join and leave the system since their sessions constitute the majority of overall peer arrival into the system.

Theorem 1. *With Assumption 1 and any finite $n \geq 1$:*

$$F(n, x) = \sum_{i=1}^{n} b_i F_i(x), \quad l(n) = \sum_{i=1}^{n} b_i l_i, \tag{9}$$

where $b_i := \lambda_i / \sum_{j=1}^{n} \lambda_j$ and λ_i is defined in (3)*.*

Proof. For large t, set $S(t)$ contains approximately

$$f_i(t) = \frac{\lfloor t\lambda_i \rfloor}{\sum_{j=1}^{n} \lfloor t\lambda_j \rfloor} \tag{10}$$

lifetime variables from user i. Bounding this metric, we have:

$$b_i - \frac{1}{\sum_{j=1}^{n} t\lambda_j} \leq f_i(t) \leq \frac{t\lambda_i}{\sum_{j=1}^{n} t\lambda_j - n}, \tag{11}$$

where $b_i = \lambda_i / \sum_{j=1}^{n} \lambda_j$. Sending t to infinity in (11), it immediately follows that the proportion of r.v.'s from user i in $S(t)$ converges to $\lim_{t\to\infty} f_i(t) = b_i$. Therefore, the probability that the value of variable in set $S(t)$ is no larger than fixed $x \geq 0$

converges to:

$$\lim_{t\to\infty} F(n,t,x) = \lim_{t\to\infty} \sum_{i=1}^{n} P(L_i \le x) f_i(t)$$
$$= \sum_{i=1}^{n} P(L_i \le x) \lim_{t\to\infty} f_i(t), \tag{12}$$

showing that the time limiting distribution exists.

Recalling that each $l_i < \infty$ by Assumption 1-b), we integrate the tail distribution $1 - F(n,x)$ for finite n to obtain:

$$E[L(n)] = \int_0^\infty \left(1 - \sum_{i=1}^{n} b_i F_i(x)\right) dx$$
$$= \sum_{i=1}^{n} b_i \int_0^\infty (1 - F_i(x)) dx,$$

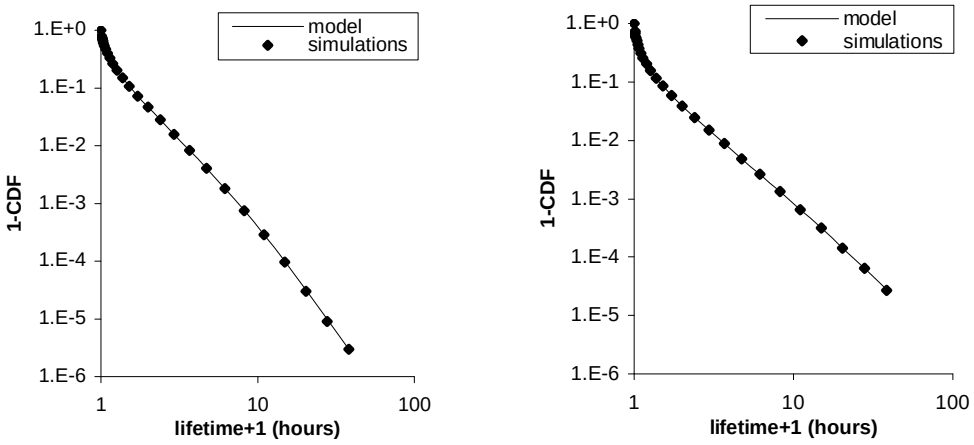
which leads to desired results in (9). □

Observe from (9) that the expected time that users stay in the system is equal to the mean system population $\sum_i \lambda_i l_i = \sum_i a_i$ divided by the aggregate user arrival rate $\sum_i \lambda_i$, which is consistent with Little's law.

Theorem 1 holds under the more general Assumption 1 as long as n is finite; however, to guarantee that the sums in (9) converge one requires Assumption 2. We show this analysis later in the chapter. In the meantime, we state similar results for aggregate offline durations.

Corollary 1. *With Assumption 1 and any finite $n \ge 1$, the CDF of aggregate offline durations is $G(n,x) := \sum_{i=1}^{n} b_i G_i(x)$ and the its mean is $d(n) := \sum_{i=1}^{n} b_i d_i$.*

We verify (9) in simulations and discuss several implications of this result. Two typical simulations are presented in Fig. 6 for exponential and heavy-tailed lifetimes, both of which show that the model is very consistent with simulation results. Both



(a) system $\mathcal{E}$ (b) system $\mathcal{H}$

Fig. 6. Comparison of simulation results of $F(n,x)$ to model (9) in a graph with $n = 1000$ nodes. System evolved to age 10^5 hours.

figures are on log-log scale and plot $1 - F(n,x)$ vs. $1 + x$ to make the shifted Pareto distribution in (2) appear as a straight line. Notice in Fig. 6(a) that system $\mathcal{E}$ produces an appearance of a heavy-tailed aggregate distribution $F(n,x)$ *even though all individual* $F_i(x)$ *are exponential*. This can be explained as follows. It is well-known [20] that for a hyper-exponential distribution in the form of (9) and *any* desired distribution $W(x)$ with a monotonic PDF (probability density function), there exists a set of weights $\{b_1, \ldots, b_n\}$ such that (9) converges to $W(x)$ as $n \to \infty$. Given numerous possibilities for the arrival-rate set $\{\lambda_1, \ldots, \lambda_n\}$ in practice, it is possible that one can observe a nicely shaped Pareto, Weibull, or other distribution $F(n,x)$, which is produced by a mixture of exponential $F_i(x)$. It may therefore be premature to conclude that Pareto $F(n,x)$ measured experimentally [12], [74] necessarily reveals the true nature of individual user behavior.

While our current conclusion shows that one cannot characterize the lifetimes or availability of individual peers by observing their aggregate behavior, the next question we seek to answer is *whether the aggregate behavior* $F(n,x)$ *can be used to*

characterize the parameters of a single user selected from the system randomly?

3.2. Characteristics of Selected Users

Suppose v picks a random currently-alive user i as a potential neighbor. Our primary goal is to understand the properties of i in terms of two metrics: its remaining online duration and its current session length.

3.2.1 Definitions

Let $R_i(t)$ denote the remaining life of a given user i at time t, i.e., the remainder of the current ON cycle illustrated in Fig. 4. Variable $R_i(t)$ is important since it determines how long this neighbor will remain online *after it has been selected*. The equilibrium residual lifetime distribution $H_i(x) := \lim_{t\to\infty} P(R_i(t) \leq x | Z_i(t) = 1)$ can be written in terms of $F_i(x)$ [91]:

$$H_i(x) = \frac{1}{l_i} \int_0^x (1 - F_i(u))du, \quad x \geq 0. \tag{13}$$

Next, define $R(n,t)$ to be the residual lifetime of the user *randomly selected* from among $N(n,t) \geq 1$ users that are alive. Denote by $H(n,x)$ the equilibrium distribution of $R(n,t)$ conditioned on $N(n,t) \geq 1$:

$$H(n,x) := \lim_{t\to\infty} P(R(n,t) \leq x | N(n,t) \geq 1). \tag{14}$$

Our goal is to obtain an expression for (14). We start with the most general case where choices may be based on the lifetimes of potential neighbors and then proceed to the much-simpler case of uniform selection.

3.2.2 General Case

To understand the results that follow, denote by

$$S_i(t) := \begin{cases} 1 & \text{user } i \text{ is selected by } v \text{ at } t \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

the indicator process that shows whether user i is randomly selected at time t from among $N(n,t) \geq 1$ users currently in the system. Define

$$\pi_i(x) = \lim_{t\to\infty} P(S_i(t) = 1 | Z_i(t) = 1, R_i(t) \leq x, N(n,t) \geq 1) \tag{16}$$

to be the equilibrium probability that user i is selected given that it is alive, its residual is no larger than x, and the system contains at least one user. We next elaborate on this metric.

In systems where the residual lifetime distribution of a user does not affect its chance of being chosen, $\pi_i(x) = \pi_i$ is not a function of x. This holds only in cases when neighbor selection is *independent* of the lifetimes (or ages) of selected users (e.g., this model was used in [43]). Examples that satisfy this condition include uniform selection, selection based on content similarity or random hashing space, age-independent popularity, etc. On the other hand, selection based on the age of potential neighbors or random walks (which depend on the in-degree of each user, which in turn depends on age) do not fall into this category (e.g., [96]).

Under uniform selection, each user i is picked with probability (conditioning on i being alive):

$$\begin{aligned} \pi_i(x) = \pi_i &= \lim_{t\to\infty} E[S_i(t) | Z_i(t) = 1, N(n,t) \geq 1] \\ &= \lim_{t\to\infty} E\Big[\frac{1}{N_n^i(t)+1}\Big], \end{aligned} \tag{17}$$

where $N_n^i(t) = \sum_{j=1,j\neq i}^{n} Z_j(t)$ is the population excluding user i.

For stationary random walks, $\pi_i(x)$ becomes the limiting version of expectation $E[d_i(t)/\sum_{m=1}^{N(n,t)} d_m(t)|Z_i(t) = 1, R_i(t) \leq x, N(n,t) \geq 1]$, where $d_i(t)$ is node degree of user i at time t. For content-based selection, assume that each user shares w_i files with others and that each peer is selected to be a neighbor proportionally to its "content utility" w_i. Then, the selection probability in (17) may be equal to $E[w_i/\sum_{m=1}^{N(n)} w_m|N(n) \geq 1]$.

As must be evident, the general model above can implement quite complex rules for choosing neighbors; however, tractability of the resulting distribution$H(n,x)$ is questionable for all except the simplest cases. Below, we first derive $H(n,x)$ for the most generic case and show that it can be expressed as a sum of weighted individual residual distributions, where the weights are biased towards users with large availability a_i and high probability $\pi_i(x)$ of being selected. We later simplify this expression for uniform selection.

Lemma 2. *Given Assumption 1 and finite $n \geq 1$:*

$$H(n,x) = \sum_{i=1}^{n} a_i \pi_i(x) H_i(x), \tag{18}$$

where $\pi_i(x)$ is given by (16)*.*

Proof. Recalling the additivity rule for disjoint events, define $q_i(x,t) = P(R_i(t) \leq x, S_i(t) = 1, Z_i(t) = 1|N(n,t) \geq 1)$ and re-write (14) as $H(n,x) = \lim_{t\to\infty} \sum_{i=1}^{n} q_i(x,t)$. For ease of presentation, break $q_i(x,t)$ into a product of the following two terms using conditional probabilities:

$$\begin{aligned} a(x,t) &= P(S_i(t) = 1|Z_i(t) = 1, R_i(t) \leq x, N(n,t) \geq 1) \\ b(x,t) &= P(Z_i(t) = 1, R_i(t) \leq x|N(n,t) \geq 1) \end{aligned} \tag{19}$$

It is now easy to notice that $\lim_{t\to\infty} a(x,t) = \pi_i(x)$ and $\lim_{t\to\infty} b(x,t) = a_i H_i(x)$, which leads to (18). □

Next, we focus on $H(n,x)$ under uniform selection and leave analysis of other strategies to future work.

3.2.3 Uniform Selection

While (18) under uniform selection has a simpler shape

$$H(n,x) = \sum_{i=1}^{n} a_i \pi_i H_i(x), \tag{20}$$

the expectation in π_i remains to be expanded in closed-form. Our first auxiliary result establishes important properties of $E[1/N(n)|N(n) \geq 1]$.

Lemma 3. *Given Assumption 2 and $N(n) \geq 1$, $\mu_n/N(n)$ converges to 1 in r-th mean for all $r \geq 1$:*

$$\lim_{n\to\infty} E\left[\left|\frac{\mu_n}{N(n)} - 1\right|^r \; |N(n) \geq 1\right] = 0, \tag{21}$$

where $\mu_n = E[N(n)]$ is the mean population.

Proof. Define $A_n := N(n)/\mu_n$, *given* $N(n) \geq 1$. In what follows, we first prove that $A_n \xrightarrow{p} 1$ (i.e., convergence in probability), then that $A_n^{-1} \xrightarrow{p} 1$, and finally show uniform integrability [10] of A_n^{-r} for constant $r \geq 1$.

Chebyshev's inequality implies

$$\forall \epsilon > 0, \quad P\left(\left|\frac{N(n)}{\mu_n} - 1\right| \geq \epsilon\right) \leq \frac{Var[N(n)]}{\epsilon^2 \mu_n^2} \to 0, \tag{22}$$

as $n \to \infty$, since $\mu_n = \Theta(n)$ and $Var[N(n)] = \sum_i a_i(1-a_i) = \Theta(n)$ from Lemma 1. Meanwhile, applying the Chernoff bound for the sum of independent Bernoulli

variables $N(n)$, we have that for any constant $c > 0$,

$$P(N(n) \geq c) \geq 1 - \exp(-\mu_n(1 - c\mu_n^{-1})^2/2) \to 1, \tag{23}$$

as $n \to \infty$. It follows from (22)-(23) that

$$\begin{aligned} \forall \epsilon > 0,\ P(|A_n - 1| \geq \epsilon) &= P\left(\left|\frac{N(n)}{\mu_n} - 1\right| \geq \epsilon | N(n) \geq 1\right) \\ &\leq P\left(\left|\frac{N(n)}{\mu_n} - 1\right| \geq \epsilon\right) / P(N(n) \geq 1) \to 0, \end{aligned} \tag{24}$$

as $n \to \infty$. The above shows that $A_n \xrightarrow{p} 1$ as $n \to \infty$.

Next, note that $g(x) := 1/x$ is a continuous function for all $x > 0$. Since $1/A_n > 0$ given $N(n) \geq 1$, using (24) and the continuity theorem [10, pp. 112] lead to

$$\lim_{n\to\infty} P\left(\left|A_n^{-1} - 1\right| \geq \epsilon\right) = 0, \tag{25}$$

indicating that $A_n^{-1} \xrightarrow{p} 1$ in the limit.

Our final step is to show that the following condition holds in order to prove uniform integrability of A_n^{-r}:

$$\sup_n E\left[|A_n^{-r}| \mathbf{1}_{|A_n^{-r}|>\alpha}\right] \to 0, \tag{26}$$

as $\alpha \to \infty$. To this end, note that given $N(n) \geq 1$, we have $A_n^{-r} \leq \mu_n^r \leq n^r$, $r \geq 1$. It is thus clear that for $n < \alpha^{1/r}$, $E[|A_n^{-r}| \mathbf{1}_{|A_n^{-r}|>\alpha}] = 0$. This leads to

$$\begin{aligned} \sup_n E\left[|A_n^{-r}| \mathbf{1}_{|A_n^{-r}|>\alpha}\right] &= \sup_{n \geq \alpha^{1/r}} E\left[|A_n^{-r}| \mathbf{1}_{|A_n^{-r}|>\alpha}\right] \\ &\leq \sup_{n \geq \alpha^{1/r}} \mu_n^r E\left[\mathbf{1}_{|A_n^{-r}|>\alpha}\right], \end{aligned} \tag{27}$$

where $E[\mathbf{1}_{|A_n^{-r}|>\alpha}] = P(|A_n^{-r}| > \alpha)$ will be examined next.

By the Chernoof bound, we have that for *all* $n \geq 1$,

$$\begin{aligned} P(|A_n^{-r}| > \alpha) &= P(N(n) < \alpha^{-1/r}\mu_n \ | N(n) \geq 1) \\ &\leq P(N(n) < \alpha^{-1/r}\mu_n)/P(N(n) \geq 1) \\ &\leq \frac{\exp\Big(-\mu_n(1-\alpha^{-1/r})^2/2\Big)}{1-\exp(-\mu_n(1-\mu_n^{-1})^2/2)}. \end{aligned} \tag{28}$$

Using the upper bound in (28) and noting that for $n \geq \alpha^{1/r}$, $\mu_n \to \infty$ as $\alpha \to \infty$, (27) yields

$$\sup_n E\left[|A_n^{-r}|\mathbf{1}_{|A_n^{-r}|>\alpha}\right] \leq \sup_{n\geq\alpha^{1/r}} \mu_n^r P(|A_n^{-r}| > \alpha) \to 0,$$

as $\alpha \to \infty$, which proves that (26) holds.

Equipped with (25) and (26), applying Theorem 5 in [10, pp. 113] immediately establishes this lemma. □

Invoking Lemma 4, we readily obtain the following result.

Lemma 4. *Given Assumption 2, $N(n) \geq 1$, and constant c, we have that for all $r \geq 1$,*

$$\lim_{n\to\infty} E\left[\Big(\frac{\mu_n}{N(n)+c}\Big)^r \ | N(n) \geq \max(1, 1-c)\right] = 1. \tag{29}$$

Proof. Note from (23) that $N(n) \geq \max(1, 1-c)$ holds w.p. 1 as $n \to \infty$. This allows us to replace the condition in (21) with $N(n) \geq \max(1, 1-c)$ to reach

$$a_n := E\Big[\Big|\frac{\mu_n}{N(n)} - 1\Big|^r \ | N(n) \geq \max(1, 1-c)]\Big] \to 0, \tag{30}$$

as $n \to \infty$. It is then clear that $\mu_n^r E[1/N^r(n)|N(n) \geq \max(1, 1-c)] = 1$. This

directly leads to

$$b_n := E\left[\left|\frac{\mu_n}{N(n)+c} - \frac{\mu_n}{N(n)}\right|^r \ \middle| N(n) \geq \max(1, 1-c)\right]$$
$$= \Theta(\mu_n^{-r}) \to 0, \tag{31}$$

as $n \to \infty$. Further, since $|f+g|^r \leq 2^r(|f|^r + |g|^r)$ for $r \geq 1$,

$$\lim_{n\to\infty} E\left[\left|\frac{\mu_n}{N(n)+c} - 1\right|^r \ \middle| N(n) \geq \max(1, 1-c)\right]$$
$$\leq \lim_{n\to\infty} 2^r(b_n + a_n) = 0, \tag{32}$$

where the last step is obtained using (30) and (31).

Finally, the convergence in r-th mean shown in (32) immediately leads to (29) by Minkowski's inequality. □

In order to tackle the convergence of the sum in (20), our second auxiliary result shows that both $F(n,x)$ and $l(n)$ have limiting distributions.

Lemma 5. *Under Assumption 2, the following sequences converge almost surely (a.s.) as $n \to \infty$:*

$$F(n,x) \xrightarrow{a.s.} F(x) := \frac{\sum_{j=1}^{\mathcal{T}} p_j \lambda^{(j)} F^{(j)}(x)}{\sum_{j=1}^{\mathcal{T}} p_j \lambda^{(j)}}, \tag{33}$$

$$l(n) \xrightarrow{a.s.} l := \frac{\sum_{j=1}^{\mathcal{T}} p_j a^{(j)}}{\sum_{j=1}^{\mathcal{T}} p_j \lambda^{(j)}}, \tag{34}$$

where $\lambda^{(j)} := 1/(l^{(j)} + d^{(j)})$ and $a^{(j)} := l^{(j)}/(l^{(j)} + d^{(j)})$. Furthermore, $F(x)$ is a proper CDF function and $0 < l < \infty$.

Proof. Re-writing (9), we get

$$F(n,x) = \frac{\sum_{i=1}^{n} \lambda_i F_i(x)}{n} \cdot \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \lambda_i}.$$

Since $\{\lambda_i\}$, $\{F_i(x)\}$ are i.i.d. sequences under Assumption 2, both sample means $\frac{1}{n}\sum_{i=1}^{n}\lambda_i F_i(x)$ and $\frac{1}{n}\sum_{i=1}^{n}\lambda_i$ converge as $n\to\infty$ to their expected values w.p. 1 by the strong law of large numbers, which leads to (33). Using the same reasoning for $l(n)$, we obtain (34) and complete the proof. □

Combining the last two lemmas, we have our main result.

Theorem 2. *Given Assumption 2, $H(n,x)$ converges almost surely (a.s.) to the following as $n\to\infty$:*

$$H(n,x) \xrightarrow{a.s.} H(x) := \frac{1}{l}\int_0^x (1-F(u))\,du, \tag{35}$$

where $F(x)$ and l are given in (33)-(34)*.*

Proof. Transform (20) into:

$$H(n,x) = \sum_{i=1}^{n} \frac{a_i H_i(x)}{n}\cdot n\pi_i. \tag{36}$$

We start with $n\pi_i$. Observing that

$$E\Big[\frac{\mu_n}{N(n)+1}|N(n)\geq 1\Big] \leq \mu_n\pi_i \leq E\Big[\frac{\mu_n}{N(n)}|N(n)\geq 1\Big]$$

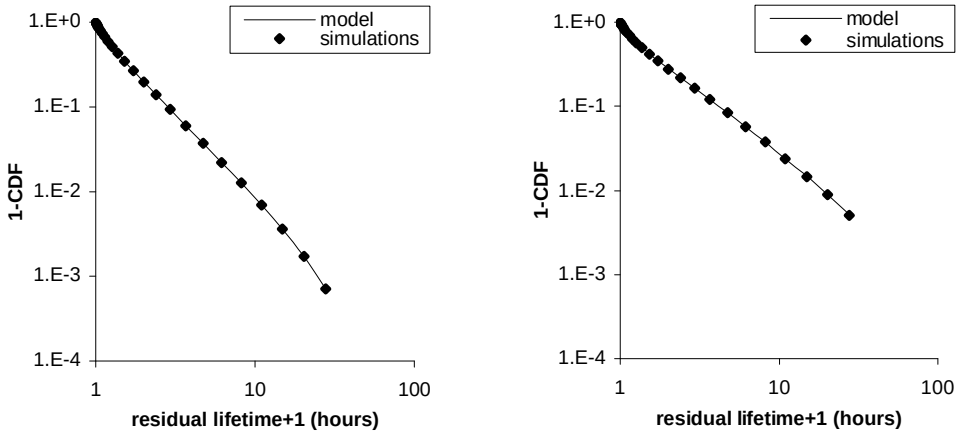
and applying Lemma 4 to both bounds, we have

$$\lim_{n\to\infty} n\pi_i = \lim_{n\to\infty}\frac{n}{\mu_n}\cdot\mu_n\pi_i = \frac{1}{\sum_{j=1}^{\mathcal{T}} p_j a^{(j)}}, \quad a.s. \tag{37}$$

The second term in (36) simplifies to:

$$\begin{aligned}\sum_{i=1}^{n}\frac{a_i H_i(x)}{n} &= \frac{\sum_{j=1}^{n}\lambda_j}{n}\sum_{i=1}^{n}\Big[\frac{\lambda_i}{\sum_{j=1}^{n}\lambda_j}\int_0^x (1-F_i(u))du\Big] \\ &\xrightarrow{a.s.} \sum_{j=1}^{\mathcal{T}}\big[p_j\lambda^{(j)}\big]\int_0^x (1-F(u))du. \end{aligned}\tag{38}$$

Combining the pieces and noticing the emergence of $1/l$, we establish (35). □



(a) system $\mathcal{E}$ (b) system $\mathcal{H}$

Fig. 7. Comparison of simulation results of $H(n,x)$ to model (39) in a graph with $n = 1000$ nodes. System age 500 hours and 10^5 iterations.

Leveraging this theorem allows us to use the following approximation:

$$H(n,x) \approx \frac{1}{l(n)} \int_0^x (1 - F(n,u))\, du = \frac{\sum_{i=1}^n a_i H_i(x)}{\sum_{i=1}^n a_i}, \tag{39}$$

which we next examine in simulations with relatively small networks. As shown in Fig. 7 for the exponential and Pareto cases, simulation results of $H(n,x)$ match the model very well and also demonstrate that $\mathcal{E}$ may produce residual lifetime distributions that appear to be non-exponential. In practice, $n \geq 50$ is often sufficient to keep (39) very accurate (simulations omitted for brevity).

Note that (35) is very important since it shows that in practice one only needs to measure the aggregate lifetime distribution $F(x)$ and its mean l rather than each $F_i(x)$ and each user availability a_i in order to obtain the residual lifetime distribution of a uniformly selected neighbor. Assuming from measurement studies [12], [30], [47], that $F(x)$ is Pareto with $F(x) = 1 - (1 + x/\beta)^{-\alpha}$, (35) reduces to:

$$H(x) = 1 - (1 + x/\beta)^{-(\alpha-1)}. \tag{40}$$

Comparing (40) to $F(x)$, we see that residuals are stochastically larger than user lifetimes, which implies that a uniformly selected user is more reliable than new arrivals in terms of failure. For other neighbor selection strategies, it is important to realize that the distribution of residual lifetimes may be completely different from (35) and should be analyzed accordingly.

3.2.4 Lifetime of Users in the System

Denote by $J(n,x)$ the equilibrium lifetime distribution of users *currently* in the system conditioned on $N(n,t) \geq 1$. As observed in [81], distribution $J(n,x)$ is clearly different from $F(n,x)$; however, no closed-form analysis has been made available to date. The intuitional rationale behind this difference is that lifetimes of the peers observed in the system are biased towards larger values, which is commonly known as the *inspection paradox* [91]. Below, we formally derive $J(n,x)$ is as a simple function of $F(n,x)$ for $n \to \infty$.

Denote by $J_i(x) := \left(xF_i(x) - \int_0^x F_i(u)du\right)/l_i$ the CDF of the current ON cycle of user i given that it is "inspected" at $t \gg 0$, i.e., its spread [91]. Since $J(n,x)$ is the same as the lifetime distribution of a uniformly randomly selected user from the set of live peers, we reach the next result following the analysis in Theorem 2.

Corollary 2. *Given Assumption 2, the lifetime distribution $J(n,x)$ of living users converges a.s. as $n \to \infty$:*

$$J(n,x) \xrightarrow{a.s.} J(x) := \frac{1}{l}\left(xF(x) - \int_0^x F(u)du\right), \tag{41}$$

where all parameters are the same as in Theorem 2.

The accuracy of (41) for finite n was confirmed in simulations, but is omitted here for brevity. Exponential lifetimes $F(x)$ imply that $J(x)$ is the Erlang(2) distribution

with mean $2E[L]$. For Pareto $F(x)$, spread $J(x)$ has no closed-form expression, but is clearly more heavy-tailed than $F(x)$. The next result summarizes these observations, as well as those of [81], in more formal terms.

Corollary 3. *With Assumption 2, spread distribution $J(x)$ is stochastically larger than $F(x)$ and the mean lifetime of a user currently alive in the system is double the mean residual lifetime of a uniformly selected user.*

3.3. Summary

This chapter introduced a simple model of churn and developed numerous closed-form results describing the behavior of users including their joint and residual lifetime distributions, evolution of system size. Our results demonstrate that given heterogeneous users and uniform selection of neighbors, both metrics $H(x)$ and $J(x)$ can be reduced to the aggregate behavior $F(x)$ of joining users as long as $n \gg 1$. The rest of the dissertation shows that $F(x)$ in such systems can be additionally used to obtain the distribution of in-degree as a function of users' age and thus completely characterize local resilience of unstructured P2P networks.

CHAPTER IV

NODE OUT-DEGREE AND AGE-BASED NEIGHBOR SELECTION*

4.1. Introduction

Traditional analysis of node isolation [42], [45] focuses on the effect of average neighbor-replacement delay $E[S]$, average user lifetime $E[L]$, and fixed out-degree k on the resilience of the system. These results show that probability ϕ with which each arriving user is isolated from the system during its lifetime is proportional to $k\rho(1+\rho)^{-k}$, where $\rho = E[L]/E[S]$. While this result is asymptotically exact under *exponential* user lifetimes and *uniform* neighbor selection, it remains to be investigated whether stronger results can be obtained for heavy-tailed lifetimes observed in real P2P networks [12], [89] and/or non-uniform neighbor selection. We study these questions below.

4.1.1 Chapter Structure and Contributions

The main focus of this chapter is to understand node isolation in the context of unstructured networks (such as Gnutella) where neighbor selection is not constrained by fixed rules. As in [42], we assume that each arriving user is assigned a random lifetime L drawn from some distribution $F(x)$ and is given k initial neighbors randomly selected from the system. The user then constantly monitors and replaces its

*Reprinted with permission from "Node Isolation Model and Age-Based Neighbor Selection in Unstructured P2P Networks," Z. Yao, X. Wang, D. Leonard, and D. Loguinov, 2009. *IEEE/ACM Transactions on Networking*, vol. 17, no. 1, pp 144-157, Copyright 2009 by IEEE.

neighbors to avoid isolation from the rest of the system. Random replacement delay S is needed to detect the failure of an old neighbor and find a new one from among the remaining alive users. Unlike [42], we allow L to come from any completely monotone distribution (a PDF $f(x)$ is *completely monotone* if derivatives $f^{(n)}$ of all orders exist and $(-1)^n f^{(n)}(x) \geq 0$ for all $x > 0$ and $n \geq 1$ [21, page 415]), e.g., Pareto and Weibull, as long as $E[L] < \infty$, and neighbor selection to be arbitrary, as long as the stationary distribution $H(x)$ of residual lifetimes R of selected neighbors is known.

We first build a generic isolation model that allows computation of ϕ with arbitrary accuracy for any completely monotone density function of residual lifetimes R. This result is achieved by replacing the distribution $H(x)$ of R with a hyper-exponential distribution, which can be performed with any accuracy, and then solving the resulting Markov chain for the probability of absorption into the isolation state before the user decides to leave the system. While this model only admits a numerical solution through matrix manipulation, it allows very accurate computation of ϕ for very heavy-tailed cases when the exponential upper bound $\phi \leq k\rho(1+\rho)^{-k}$ [42] is rather loose. The model is also necessary for studying isolation behavior of the various neighbor-selection strategies examined in later parts of the chapter where simulations are impractical or impossible due to the small values of ϕ.

The second part of the chapter verifies the model of ϕ under uniform neighbor replacement and analyzes its performance for very heavy-tailed lifetimes (i.e., $Var[L] = \infty$). We show that as the age $\mathcal{T}$ of the system becomes infinite and shape parameter α of Pareto user lifetime distribution approaches 1, the isolation probability decays to zero proportionally to $(\alpha - 1)^k$, which holds for *any* number of neighbors $k \geq 1$ and *any* search delay S, implying that such systems may achieve arbitrary resilience without replacing any neighbors. In practice, however, α is a fixed number bounded away from 1 (common studies suggest that α is between 1.06 [12] and

1.09 [89]) and $\mathcal{T}$ is finite, which cannot guarantee high levels of robustness without neighbor replacement.

As an improvement over the uniform case, we next study the so-called *max-age* neighbor selection [12], [41], [77], in which a user samples m uniformly random peers per link it creates and selects the one with the largest current age to be its neighbor. We show that larger values of m lead to stochastically larger R and improve the expected remaining lifetimes of found neighbors by a factor approximately proportional to $m^{1/(\alpha-1)}$ for $m > 1$. For example, $\alpha = 3$ increases $E[R]$ as $\sqrt{m}$, $\alpha \approx 2$ increases $E[R]$ linearly in m, and $\alpha < 2$ results in $E[R] = \infty$ regardless of m as long as $\mathcal{T} = \infty$. We do not obtain a closed-form factor of reduction for ϕ compared to the purely uniform case, but note that it is a certain monotonic function of m. This does not change, however, the qualitative behavior of ϕ under the no-replacement policy and still requires $\alpha \to 1$ to achieve $\phi \to 0$ for any fixed m.

While the max-age approach is viable and very effective in general, it relies on the system's ability to sample m peers uniformly randomly per created link. This can be accomplished using Metropolis-style random walks [99]; however, this method requires overhead that is linear in m and thus may not scale well for large m. To build a distributed solution that requires only *one* sample per link, the last part of the chapter proposes a novel technique based on random walks over directed graphs, in which the weight of in-degree edges at each node is kept proportional to the age of the corresponding user. Under these conditions, we derive a model for the residual distribution $H(x)$ and show that isolation probability ϕ converges to 0 for any $1 < \alpha \leq 2$ as system size $n \to \infty$ and age $\mathcal{T} \to \infty$, which holds for any number of neighbors $k \geq 1$ and any search delay S. Compared to the uniform and max-age cases, this is a much stronger result that shows that with just $k = 1$ neighbor and no replacement of failing neighbors, large P2P systems with $\alpha \leq 2$ can guarantee arbitrarily low values

of ϕ. We finish the chapter by studying in simulations the approach rate of ϕ to 0 and its effect in practice.

4.2. General Node Isolation Model

In this section, we build a model for the probability ϕ that a node v becomes isolated due to all of its neighbors simultaneously reaching the failed state during its lifetime.

4.2.1 Background

We assume that user join/departure processes follow the user churn model in Chapter III. For neighbor dynamics, we adopt conventions of [43]. Upon joins, user v finds k initial neighbors and then continuously monitors their presence in the system. Neighbor replacement occurs only when an existing neighbor fails. Each neighbor i is either alive (i.e., ON) or dead (i.e., OFF) at any time t. The random ON duration R is the residual lifetime of the neighbor from the instance it is selected by v until its departure. The random OFF duration S is search delay until a replacement is found. Note that residuals R depend on the neighbor-selection strategy [93] and should be analyzed accordingly.

Let L be the lifetime of joining user v, drawn from the aggregate user lifetime distribution $F(x)$ that is known to our analysis (e.g., through an external measurement process [12], [89]). Further, denote by $X(t)$ the number of neighbors of user v at time t. We can then define the first-hitting time T onto the isolation state $X(t) = 0$ as:

$$T = \inf(t > 0 : X(t) = 0 | X(0) = k). \tag{42}$$

Note that T specifies the duration before user v becomes isolated (i.e., loses all of

its neighbors). The goal of this section is to derive the node isolation probability $\phi = P(T < L)$, which is the likelihood of v becoming isolated before it voluntarily decides to leave the system. For systems with non-exponential user lifetimes, out-degree process $\{X(t)\}$ is not Markovian, which makes closed-form derivation of ϕ very difficult. However, certain cases identified below can be solved with arbitrary accuracy by replacing residual lifetimes and search delays with their hyper-exponential equivalents.

The rest of this section deals with constructing a continuous-time Markov chain that keeps track of v's out-degree under the hyper-exponential approximation and leads to very accurate closed-form models of T and ϕ.

4.2.2 Hyper-Exponential Approximation

Recall that the hyper-exponential distribution H_m is a mixture of m exponential random variables with probability density function (PDF) in the form of [91]:

$$f_H(x) = \sum_{j=1}^{m} p_j \mu_j e^{-\mu_j x}, \tag{43}$$

where $\mu_j, p_j \geq 0$ for all j and $\sum_{j=1}^{m} p_j = 1$. The above distribution can be interpreted as generating each exponential random variable $\exp(\mu_j)$ with probability p_j. It is well-known [20] that any *completely monotone* density function $f(x)$ can be represented with any desired accuracy using (43), i.e., $f_H(x) \to f(x)$ as $m \to \infty$. In the analysis below, we leverage this property of hyper-exponentials and the fact that Pareto and Weibull residual PDFs are completely monotone. While some of the prior literature [20] has used as many as 14 exponentials to approximate Pareto $f(x)$, our analysis suggests that as few as 3 are usually sufficient for achieving very accurate results on ϕ (see below).

Before we proceed with the derivations, it is useful to visualize the meaning of

hyper-exponential distributions in our lifetime model. Given that the PDF of neighbor residual lifetimes R is $f_R(t) = \sum_{i=1}^{r} p_i \mu_i e^{-\mu_i t}$, imagine that there are r different types of neighbors, where residual lifetimes of peers of type i are exponentially distributed with rate μ_i for $i = 1, \ldots, r$. When v requires a new neighbor, it selects a node of type i with probability p_i. Similarly, provided that the PDF of search delay S is $f_S(t) = \sum_{j=1}^{s} q_i \lambda_j e^{-\lambda_j t}$, suppose that there are s types of searches that can be currently in progress. A search of type j is instantiated by v with probability q_j and has duration exponentially distributed with rate λ_j for $j = 1, \ldots, s$.

Given that there are r types of neighbors and s types of search processes, define $W(t)$ to be a random process that counts the number of v's neighbors and searches of each type at time t:

$$W(t) = (X_1(t), \ldots, X_r(t), Y_1(t), \ldots, Y_s(t)), \tag{44}$$

where $X_i(t)$ is the number of v's neighbors of type i at time t for $i = 1, \ldots, r$ and $Y_j(t)$ the number of searches in progress of type j at time t for $j = 1, \ldots, s$. Also note that v's out-degree $X(t) = \sum_{i=1}^{r} X_i(t)$ is fully described by process $\{W(t)\}$. The state space Ω for $\{W(t)\}$ is:

$$\Omega = \{(x_1, ..., x_r, y_1, ..., y_s)\}, \tag{45}$$

where $x_i \in \{0, 1, \ldots, k\}$, $y_j \in \{0, 1, \ldots, k\}$, and $\sum_{i=1}^{r} x_i + \sum_{j=1}^{s} y_j = k$. As long as neighbor residual lifetimes R and search delays S can be reduced to the hyper-exponential distribution, the resulting process $\{W(t)\}$ can be viewed as a homogenous continuous-time Markov chain as we show next.

Theorem 3. *Given that the density function of residual lifetimes* $f_R(t) = \sum_{j=1}^{r} p_j \mu_j e^{-\mu_j t}$ *and the density function of search times* $f_S(t) = \sum_{j=1}^{s} q_j \lambda_j e^{-\lambda_j t}$, $\{W(t)\}$ *is a homo-*

geneous continuous-time Markov chain with a transition rate matrix Q given below.

Proof. Since neighbors of type i are $\exp(\mu_i)$ and search processes of type j are $\exp(\lambda_j)$, the sojourn time in state $u = (x_1, ..., x_r, y_1, ..., y_s)$ is exponential with rate:

$$\Lambda_u = \sum_{i=1}^{r} x_i \mu_i + \sum_{j=1}^{s} y_j \lambda_j. \tag{46}$$

Observe that when a neighbor dies, a search starts immediately and its properties are independent of those of the existing searches or neighbor lifetimes. Conversely, when a search ends and a new neighbor is found, the characteristics of this neighbor are independent of any previous behavior of $\{W(t)\}$. This independence allows us to easily write transition probabilities between adjacent states of $\{W(t)\}$.

The first type of transition reduces $W(t)$ by 1 in response to the failure of one of v's neighbors, which is equivalent to a jump from state:

$$(x_1, ..., x_i, ..., x_r, y_1, ..., y_j, ..., y_s) \tag{47}$$

to state:

$$(x_1, ..., x_i - 1, ..., x_r, y_1, ..., y_j + 1, ..., y_s) \tag{48}$$

for any suitable $x_i \geq 1$. For simplicity of notation, we call the above transition $(x_i, y_j) \to (x_i - 1, y_j + 1)$. The corresponding probability that a neighbor of type i dies and a search of type j starts is $x_i \mu_i q_j / \Lambda_u$.

The second type of transition increases $W(t)$ by 1 as a result of finding a replacement neighbor, which corresponds to a jump from state:

$$(x_1, ..., x_i, ..., x_r, y_1, ..., y_j, ..., y_s) \tag{49}$$

to state:

$$(x_1, ..., x_i + 1, ..., x_r, y_1, ..., y_j - 1, ..., y_s) \tag{50}$$

for any $y_j \geq 1$. The corresponding notation for this transition is $(x_i, y_j) \to (x_i + 1, y_j - 1)$. The related probability that a search process of type j ends and finds a new neighbor of type i before any other event happens is $y_j \lambda_j p_i / \Lambda_u$.

By recognizing that the jumps behave like a discrete-time Markov chain and the sojourn times at each state are independent exponential random variables, we immediately conclude that $\{W(t)\}$ is a homogeneous continuous-time Markov chain with a transition rate matrix $Q = (q_{uu'})$ where

$$q_{uu'} = \begin{cases} q_j x_i \mu_i & (x_i, y_j) \to (x_i - 1, y_j + 1) \\ p_i y_j \lambda_j & (x_i, y_j) \to (x_i + 1, y_j - 1) \\ -\Lambda_u & u' = u \\ 0 & \text{otherwise} \end{cases}, \tag{51}$$

are transition rates from u to u', which represent any suitable states in the form of (45) that satisfy transition requirements on the right side of (252). □

Using notation $W(t)$, the first-hitting time T in (42) can now be rewritten as:

$$T = \inf\Big(t > 0 : \sum_{i=1}^{r} X_i(t) = 0 \Big| \sum_{i=1}^{r} X_i(0) = k\Big), \tag{52}$$

where $X_i(t)$ is defined in (44). The next step is to obtain the initial state distribution of $\{W(t)\}$ and derive the PDF of the first-hitting time T based on the transition rate matrix Q in (252). For small values of k, the matrix can be easily represented in memory and manipulated in software packages such as Matlab. For example, when $r = s = 3$ commonly used in this work, the size of Q is 252×252 for $k = 5$ and 792×792 for $k = 7$.

The initial state distribution $\pi(0)$ is in form of:

$$\pi(0) = \left(\pi_{(x_1,...,x_r,y_1,...,y_s)}(0)\right), \tag{53}$$

where each entry in the vector represents the probability that the chain starts in state $(x_1, ..., x_r, y_1, ..., y_s)$ for all possible permutations of variables x_i and y_j. Note, however, that the only valid starting states are those in which the number of alive neighbors $\sum_{i=1}^{r} x_i$ is exactly k and the number of searches in progress $\sum_{j=1}^{s} y_j$ is zero.

After rather straightforward manipulations, $\pi(0)$ can be obtained as follows.

Lemma 6. *Valid starting states have initial probabilities:*

$$\pi_{(x_1,...,x_r,0,...,0)}(0) = \prod_{i=1}^{r} \binom{k - \sum_{j=1}^{i-1} x_j}{x_i} p_i^{x_i}, \tag{54}$$

and all other states have initial probability 0.

Proof. Define X_i to be a random variable representing the number of neighbors of type i for $i = 1, \ldots, r$. Then, given a valid starting state $u = (x_1, ..., x_r, 0, ..., 0)$ for $\sum_{i=1}^{r} x_i = k$, its initial probability can be described by:

$$\pi_u(0) = P(X_1 = x_1, \ldots, X_r = x_r) = \prod_{i=1}^{r-1} q_i, \tag{55}$$

where q_i is the probability that $X_i = x_i$ conditioned on all X_j for $j < i$ being equal to their corresponding x_j:

$$q_i = P\Big(X_i = x_i \Big| \bigcap_{j=1}^{i-1} X_j = x_j\Big). \tag{56}$$

Denote by:

$$B(x; k, p) = \binom{k}{x} p^x (1-p)^{k-x}, \text{ for } x = 0, 1, \ldots, k, \tag{57}$$

the binomial distribution with success probability p. Note that $P(X_1 = x_1)$ is simply

$q_1 = B(x_1; k, p_1)$. Next, it is clear that given $X_1 = x_1$ neighbors of type 1, the probability that the initial state contains $X_2 = x_2$ neighbors of type 2 is also binomial, but with success probability $p_2/(1 - p_1)$:

$$q_2 = P(X_2 = x_2 | X_1 = x_1) = B\left(x_2; k - x_1, \frac{p_2}{1 - p_1}\right). \tag{58}$$

It can be shown that the generalized version of (58) is:

$$q_i = B\left(x_i; k - \sum_{j=1}^{i-1} x_j, \frac{p_i}{1 - \sum_{j=1}^{i-1} p_j}\right), \tag{59}$$

which after substitution into (55) and some algebra, reduces (55) to (54). □

Armed with this result, we next focus our attention on deriving ϕ.

4.2.3 Isolation Probability

Recall that Ω denotes the set of all valid states (i.e., in the form of (45) and satisfying all constraints following the equation). Denote by:

$$E = \left\{(0, ..., 0, y_1, ..., y_s) : \sum_{j=1}^{s} y_j = k\right\} \tag{60}$$

the set of states with zero out-degree. Since we are only interested in the first-hitting time T to any state in E, it suffices to assume that all states in E are absorbing. Then, for each non-absorbing state $u \in \Omega \setminus E$, its transition rate to E is given by:

$$q_{uE} = \sum_{u' \in E} q_{uu'}, \tag{61}$$

where $q_{uu'}$ is the cell of matrix Q corresponding to transitions from state u to u'. We can then write Q in canonical form as:

$$Q = \begin{pmatrix} 0 & 0 \\ \mathbf{r} & Q_0 \end{pmatrix}, \tag{62}$$

where $\mathbf{r} = (q_{uE})^T$ for $u \notin E$ is a column vector representing the transition rates to the absorbing set E and $Q_0 = (q_{uu'}, u, u' \in \Omega \setminus E)$ is the rate matrix obtained by removing the rows and columns corresponding to states in E from Q. The following lemma shows that the PDF of T is fully determined by $\pi(0)$ and Q.

Lemma 7. *For residual lifetimes and search delays with hyper-exponential distributions, the PDF of T is given by:*

$$f_T(t) = \pi(0) V D(t) V^{-1} \mathbf{r}, \tag{63}$$

where $\pi(0)$ is the initial state distribution in (54), V is a matrix of eigenvectors of Q_0, $D(t) = diag(e^{\xi_j t})$ is a diagonal matrix, $\xi_j \leq 0$ is the j-th eigenvalue of Q_0, and Q_0 and $\mathbf{r}$ are in (253).

Proof. Generalize the first hitting time from a starting state $w \in \Omega \setminus E$ to any absorbing state in E as:

$$T_{wE} = \inf\{t > 0 : W(t) \in E | W(0) = w\}. \tag{64}$$

For regular Markov chains [70, p. 375], it is not difficult to see that T_{wE} has a continuous density function $f_{T_{wE}}(t)$ such that for small dt:

$$P(t < T_{wE} < t + dt) = f_{T_{wE}}(t)dt + o(dt). \tag{65}$$

At the same time, from last-step analysis [37, p. 211], [70, p. 388] we have:

$$P(t < T_{wE} < t + dt) = \sum_{u \in \Omega \setminus E} p_{wu}(t) q_{uE} dt + o(dt), \tag{66}$$

where $p_{wu}(t) = P(W(t) = u | W(0) = w)$ is the probability that the chain is in state u at time t given that it started in state w and q_{uE} is transition rate from state u to

any absorbing state in E. Combining (65)-(66) and letting $dt \to 0$, we easily obtain:

$$f_{T_{wE}}(t) = \sum_{u \in \Omega \setminus E} p_{wu}(t) q_{uE}. \tag{67}$$

Notice from the above that computation of $f_{T_{wE}}(t)$ requires transition probabilities $p_{wu}(t)$ for all $u \in \Omega \setminus E$, which are rather difficult to obtain in explicit closed-form for non-trivial Markov chains such as ours. Instead, we offer a solution that depends on spectral properties of Q_0 and a matrix representation of $p_{wu}(t)$ in the analysis that follows.

Expressing (67) in matrix form, we have:

$$(f_{T_{wE}}(t))^T = P_0(t)\mathbf{r}, \quad w \in \Omega \setminus E, \tag{68}$$

where $(f_{T_{wE}}(t))^T$ is a column vector, $P_0(t) = (p_{wu}(t))$ for $w \in \Omega \setminus E, u \in \Omega \setminus E$ are transition probability functions corresponding to non-absorbing states, and $\mathbf{r} = (q_{uE})^T$ for $u \in \Omega \setminus E$ is a transition rate column vector. Then representing $P_0(t) = e^{Q_0 t}$ using matrix exponential [70] and $Q_0 = V \Lambda V^{-1}$ using eigen-decomposition [59], where Q_0 is given in (253), we get:

$$P_0(t) = e^{Q_0 t} = V e^{\Lambda t} V^{-1} = V D(t) V^{-1}, \tag{69}$$

where $D(t) = \text{diag}(e^{\xi_j t})$, $\xi_j \leq 0$ is the j-th eigenvalue of Q_0, and V is a matrix of eigenvectors of Q_0. Substituting (69) into (68), we obtain:

$$(f_{T_{wE}}(t))^T = V D(t) V^{-1} \mathbf{r}, \quad w \notin E. \tag{70}$$

Finally, the PDF $f_T(t)$ of the first hitting time T is simply the product of row vector $\pi(0)$ and column vector $(f_{T_{wE}}(t))^T$:

$$f_T(t) = \pi(0)(f_{T_{wE}}(t))^T = \pi(0) V D(t) V^{-1} \mathbf{r}, \quad w \notin E, \tag{71}$$

where $\pi(0)$ is given by (54) for Markov chain $\{W(t)\}$. □

With Lemma 7 in hand, integrating $f_T(t)$ using the distribution of user lifetimes immediately leads to the following theorem.

Theorem 4. *For hyper-exponential residual lifetimes and search delays, the probability of isolation is:*

$$\phi = \pi(0)VBV^{-1}\mathbf{r}, \tag{72}$$

where $B = diag(b_j)$ is a diagonal matrix with:

$$b_j = \int_0^\infty (1 - F(t))e^{\xi_j t}dt, \tag{73}$$

$F(t)$ is the CDF of user lifetimes, and all other parameters are the same as in Lemma 7.

Proof. Note that for node v with lifetime L, its isolation probability is give by:

$$\begin{aligned}\phi = P(T < L) &= \int_0^\infty P(L > t) f_T(t)dt \\ &= \int_0^\infty (1 - F(t)) f_T(t)dt,\end{aligned} \tag{74}$$

where $F(t)$ is the CDF of user lifetimes. Invoking Lemma 7 and integrating $1 - F(t)$ using $f_T(t)$, we immediately obtain:

$$\phi = \pi(0)V\Big(\int_0^\infty (1 - F(t))D(t)dt\Big)V^{-1}\mathbf{r}, \tag{75}$$

which directly leads to (72). □

Using rate matrix Q_0, vector $\mathbf{r}$, and (72)-(256), the solution to node isolation probability ϕ can be easily computed using numerical packages such as Matlab. We perform this task next.

4.2.4 Verification of Isolation Model

We examine the accuracy of (72)-(256) using the simplest example of uniform selection. We first explore the exponential case for comparison purposes and then derive the same metric for Pareto lifetimes.

For exponential lifetimes, the next lemma immediately follows upon recalling that neighbor residual lifetimes R are also exponentially distributed with $m = 1$ in (43) due to the memoryless property of the distribution.

Lemma 8. *For exponential $L \sim \exp(\mu)$ and search delays with a hyper-exponential density $f_S(x)$, the transition rate matrix Q of $\{W(t)\}$ is given by (252) with $r = 1$, $p_1 = 1$, and $\mu_1 = \mu$. Isolation probability ϕ is in form of (72) where (256) is simply:*

$$b_j = 1/(\mu - \xi_j), \tag{76}$$

Proof. Due to the memoryless property of exponential distributions, it is clear that residual lifetimes R have the same distribution as user lifetimes L, i.e., $R \sim F(x)$. Thus we have $f_R(x) = \mu e^{-\mu x}$, requiring only one exponential in the hyper-exponential mixture model (43). Next, re-writing (256) using $F(t) = 1 - e^{-\mu t}$ for exponential lifetimes, we get:

$$b_j = \int_0^\infty e^{-\mu t} e^{\xi_j t} dt = \frac{1}{\mu - \xi_j}, \tag{77}$$

which combined with (72) immediately establishes this theorem. □

Our next theorem derives ϕ for Pareto lifetimes with the following CDF:

$$P(L < x) = 1 - \left(1 + \frac{x}{\beta}\right)^{-\alpha}, \tag{78}$$

for shape parameter $\alpha > 1$, scale parameter $\beta > 0$, and $x \geq 0$. Denote by R the residual lifetime of a uniformly random user in the system. Assuming a sufficiently

large system age $\mathcal{T}$, it follows from Theorem 2 in the previous chapter that the CDF of R under uniform selection is given by:

$$P(R < x) = 1 - \left(1 + \frac{x}{\beta}\right)^{-(\alpha-1)}. \tag{79}$$

It is clear from (79) that the PDF of Pareto residuals is completely monotone and thus can be fitted with its hyper-exponential equivalent. Invoking Theorem 4, we immediately obtain the following.

Lemma 9. *For Pareto $L \sim 1 - (1 + x/\beta)^{-\alpha}$ and hyper-exponential search delays, the transition rate matrix Q is shown in (252) where p_i and μ_i for $i = 1, \ldots, r$ are given by the hyper-exponential approximation of Pareto R with shape $\alpha - 1$ in (79). Isolation probability ϕ is given in (72) where (256) is:*

$$b_j = \beta e^{-\xi_j \beta} E_\alpha(-\xi_j \beta), \tag{80}$$

where $E_\alpha(x) = \int_1^\infty e^{-xu} u^{-\alpha} du$ is the generalized exponential integral.

Proof. Invoking Theorem 4 and using $F(t) = 1 - (1 + t/\beta)^{-\alpha}$, (256) yields:

$$b_j = \int_0^\infty \left(1 + \frac{t}{\beta}\right)^{-\alpha} e^{\xi_j t} dt = \beta e^{-\xi_j \beta} \int_1^\infty u^{-\alpha} e^{\xi_j \beta u} du, \tag{81}$$

which completes the proof by recognizing that:

$$E_\alpha(x) = \int_1^\infty e^{-xu} u^{-\alpha} du. \tag{82}$$

is the generalized exponential integral. □

We perform simulations to see the accuracy of analytical results in systems with *finite* age and size. To observe the accuracy of Lemmas 8-9, we run simulations over different distributions of search times on a graph with $n = 1,000$ nodes, $k = 7$, and mean lifetime $E[L] = 0.5$ hours (additional simulations produce similar results and

Table I. Comparison of model ϕ to simulations under uniform selection with $E[L] = 0.5$ hours and $k = 7$

$E[S]$	Pareto L with $\alpha = 3$						Exponential L	
hours	Pareto S with $\alpha = 3$		Weibull S with $c = 0.7$		Exponential S		Pareto S with $\alpha = 3$	
	Simulations	Model (80)	Simulations	Model (80)	Simulations	Model (80)	Simulations	Model (76)
.001		1.11×10^{-16}		1.12×10^{-16}		1.12×10^{-16}		4.40×10^{-16}
.01		8.49×10^{-11}		8.45×10^{-11}		9.05×10^{-11}		3.70×10^{-10}
.05	4.56×10^{-7}	4.49×10^{-7}	4.93×10^{-7}	4.96×10^{-7}	6.27×10^{-7}	6.28×10^{-7}	2.31×10^{-6}	2.31×10^{-6}
.1	1.13×10^{-5}	1.14×10^{-5}	1.21×10^{-5}	1.25×10^{-5}	1.75×10^{-5}	1.74×10^{-5}	6.01×10^{-5}	6.04×10^{-5}
.4	1.64×10^{-3}	1.64×10^{-3}	1.60×10^{-3}	1.58×10^{-3}	2.57×10^{-3}	2.59×10^{-3}	6.80×10^{-3}	6.78×10^{-3}
.6	4.43×10^{-3}	4.44×10^{-3}	4.17×10^{-3}	4.11×10^{-3}	6.67×10^{-3}	6.66×10^{-3}	1.61×10^{-2}	1.60×10^{-2}
.8	7.78×10^{-3}	7.78×10^{-3}	7.14×10^{-3}	7.16×10^{-3}	1.12×10^{-2}	1.12×10^{-2}	2.56×10^{-2}	2.56×10^{-2}

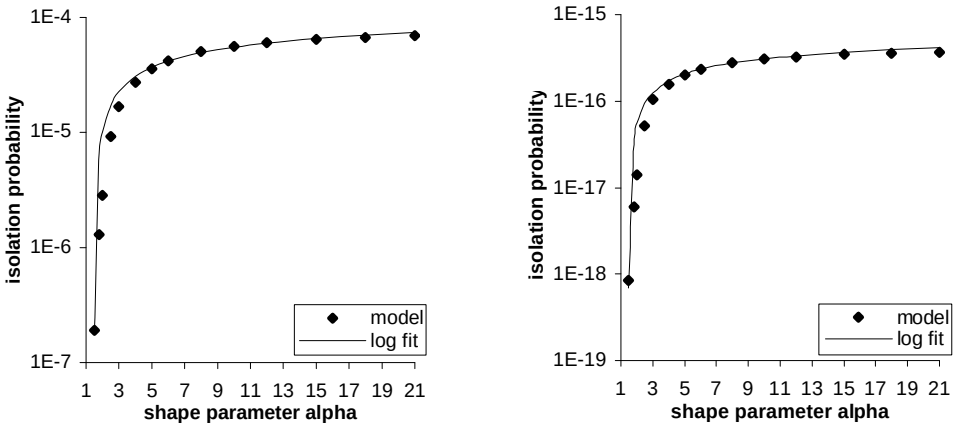
are omitted for brevity). The first search time distribution is Pareto with $\alpha = 3$ and $\beta = E[S](\alpha - 1)$ to keep the mean equal to $E[S]$. The second distribution is Weibull with CDF $1 - e^{-(x/a)^c}$ and mean $E[S] = a\Gamma(1+1/c)$. The third is exponential with rate $1/E[S]$. To compute the model, Pareto residual lifetime R is fitted with a hyper-exponential mixture model (43) using $r = 3$ and each non-exponential search distribution is fitted with model (43) using $s = 3$.

Exponential and Pareto models of ϕ are compared to simulation results in Table I. Notice in the table that both (76) and (80) are indeed very accurate for all examined search and lifetime distributions. The table also confirms that as $E[S] \to 0$, metric ϕ becomes insensitive to the distribution of S, which was earlier observed in [42] but never verified.

To understand the influence of tail weight of the lifetime distribution $F(x)$ on isolation, we use (80) to compute ϕ for several values of shape parameter α and keep $\beta = (\alpha - 1)E[L]$ to ensure that the mean lifetime $E[L]$ remains fixed. The result is shown in Fig. 8 for two values of $E[S]$ and $k = 7$. Notice in both sub-figures that the relationship between ϕ and α is similar and that ϕ appears to be approximately a logarithmic function of α for $\alpha \leq 21$, confirming that the more heavy-tailed the lifetime distribution, the smaller ϕ.

4.2.5 Necessity of Neighbor Replacement

Fig. 8 suggests that ϕ tends to 0 as α approaches 1 from above, but it is not clear at what rate this convergence takes place and whether this is indeed true. Furthermore, since $E[R] = \infty$ for $\alpha \leq 2$, a natural question arises about whether a finite system of n users and finite age $\mathcal{T}$ can in fact exhibit infinite expected residuals or $\phi = 0$ when $\alpha = 1$. We answer these questions next and show that condition $\alpha \to 1$ indeed guarantees $\phi \to 0$ even in cases when no replacement of failed neighbors is performed;



(a) $E[S] = 6$ minutes (b) $E[S] = 3.6$ seconds

Fig. 8. Impact of shape parameter α on model ϕ under uniform selection, Pareto lifetimes, $E[L] = 0.5$ hours, $\beta = (\alpha - 1)E[L]$, exponential search delays, and $k = 7$.

however, it requires that the system be *in equilibrium* (i.e., the first renewal cycle of each user must be drawn from its residual distribution or system age $\mathcal{T}$ be infinite. See [91, page 65] for a definition) by the time it is observed by an arriving user.

Theorem 5. *For an equilibrium system, Pareto lifetimes with $\alpha > 1$, and infinitely large search delays (i.e., $S = \infty$), the isolation probability is:*

$$\phi = \frac{k!}{(\gamma + 1) \times \ldots \times (\gamma + k)}, \tag{83}$$

where $\gamma = \alpha/(\alpha - 1)$. For fixed k and $\alpha \to 1$ (i.e., $\gamma \to \infty$), (83) converges to zero as $\Theta(\gamma^{-k})$.

Proof. Assuming that search delays S are infinity, the first hitting time T defined in (52) equals the maximum residual lifetime among all neighbors:

$$T = \max\{R_1, ..., R_k\}. \tag{84}$$

Then, due to the independence among k neighbors, it is easy to see that the distri-

bution of T for Pareto lifetimes under uniform selection is:

$$P(T < x) = [P(R < x)]^k = \left[1 - \left(1 + \frac{x}{\beta}\right)^{-\alpha+1}\right]^k. \tag{85}$$

It follows that given that $S = \infty$, node isolation probability is simply [42]:

$$\phi = \int_0^\infty P(T < x) f(x) dx = \frac{\Gamma(1+\gamma)k!}{\Gamma(k+1+\gamma)}, \tag{86}$$

where $f(x) = \alpha(1 + x/\beta)^{-\alpha-1}/\beta$ is the PDF of Pareto lifetimes, $\gamma = \alpha/(\alpha - 1)$, and $\Gamma(x)$ is the gamma function.

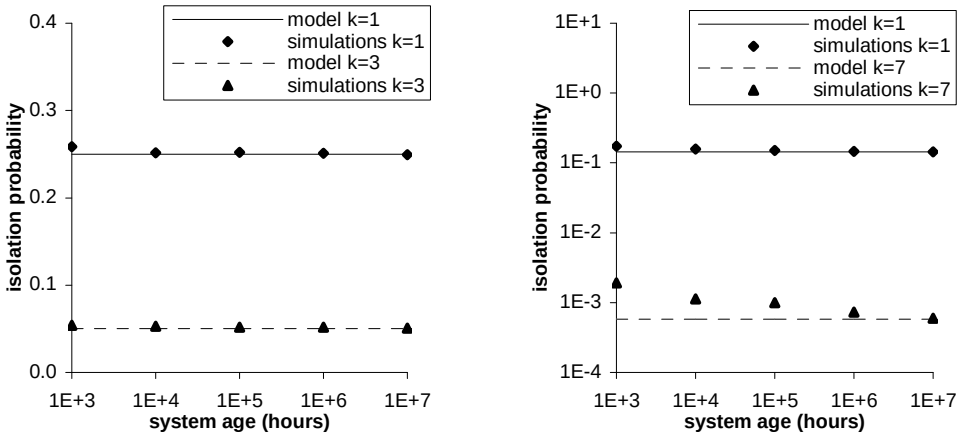
Recalling that $\Gamma(x) = (x-1)\Gamma(x-1)$ and canceling the common divisor $\Gamma(1+\gamma)$, (86) reduces to:

$$\phi = \frac{k!}{(\gamma+1) \times \ldots \times (\gamma+k)}. \tag{87}$$

As $\alpha \to 1$, it is clear that $\gamma \to \infty$, which makes ϕ in (87) converge to 0. Noticing that k is fixed, it is easy to see from (87) that $\phi = \Theta(\gamma^{-k})$. □

This result is very interesting since most prior work [42] does not consider $\alpha \leq 2$ as such cases result in infinite expected residual lifetimes, which cannot be observed in any finite system. However, if the age of the system tends to infinity, i.e., $\mathcal{T} \to \infty$, or the first lifetime of each user is drawn from the residual distribution (79), the asymptotic bound in (83) is actually achievable. In such cases, as α tends to 1, the isolation probability will decay to zero proportionally to $(\alpha - 1)^k$ as given by Theorem 5 and the system will attain any desired level of resilience without replacing neighbors. On the other hand, for α sufficiently larger than 2 studied in prior work [42], age $\mathcal{T}$ must simply exceed the convergence time to equilibrium of the underlying user-lifetime renewal process, which usually happens very quickly.

Fig. 9 shows simulation results of ϕ with $S = \infty$ and two cases of very heavy-tailed L. Notice in Fig 9(a) that for $\alpha = 1.5$, simulation results converge to model ϕ



(a) $\alpha = 1.5, S = \infty$ (b) $\alpha = 1.2, S = \infty$

Fig. 9. Convergence of simulation results to model ϕ in (83) as system age $\mathcal{T} \to \infty$ under uniform selection, no neighbor replacement, and Pareto lifetimes with $\beta = (\alpha - 1)E[L]$ in a graph with $n = 1,000$ nodes.

before system age reaches 10^4 hours (i.e., 1.14 years). However, as α reduces to 1.2, the convergence takes a much longer time as shown in Fig 9(b), where simulations approach the model when system age grows to more than $\mathcal{T} = 10^6$ hours = 114 years.

The above analysis shows that the asymptotic result $\phi \to 0$ as $\alpha \to 1$ is not readily achievable in finite P2P systems. Furthermore, recent measurement studies of user lifetimes suggest that P2P networks exhibit α that is bounded away from 1 (i.e., α is between 1.06 [12] and 1.09 [89]). Hence, most current P2P systems are not likely to satisfy the condition for $\phi \to 0$ under uniform selection and thus need to utilize either a large number of neighbors k or perform dynamic replacement of dead links with $E[S] < \infty$.

4.2.6 Discussion

While the general form of ϕ in the exact model (72) is very complex, a simple qualitative rule of increasing resilience (i.e., reducing ϕ) can be formulated based on the

properties of residual lifetimes selected by the users of a P2P system. Notice that for a fixed lifetime distribution $F(x)$, higher resilience is achieved by selecting neighbors that exhibit larger (in some sense) remaining lifetimes. Thus, given two strategies $\mathcal{S}_1$ and $\mathcal{S}_2$ for selecting neighbors, the strategy that obtains a neighbor with a larger residual lifetime during *every* replacement instance τ guarantees a lower isolation probability since the chosen neighbors survive longer and increase the chance that the current user will depart before becoming isolated. Since comparison of residual lifetimes of obtained neighbors in $\mathcal{S}_1$ and $\mathcal{S}_2$ can be performed only in the *probabilistic* sense, the above discussion can be formalized as following:

Note, however, that future residual lifetimes of sampled peers are usually not available in practice. Instead, assuming that $F(x)$ is not memoryless (i.e., non-exponential), current user age A may be used as a robust predictor of R. To understand this correlation for Pareto $F(x)$ shown in (78), consider the probability that a peer's remaining lifetime is larger than $y \geq 0$ given that its current age A is $x \geq 0$:

$$P(R > y | A = x) = \left(1 + \frac{y}{\beta + x}\right)^{-\alpha}. \tag{88}$$

Observe that the above conditional probability is a monotonically increasing function of age, i.e., the larger x, the more likely a node is to survive at least y time units in the future. This implies that *users with larger age demonstrate stochastically larger residual lifetimes R.*

This result can be generalized to all heavy-tailed distributions (defined in terms of conditional mean exceedance [32] or tail-decay rate [85], e.g., Pareto, Weibull, and Cauchy), in which the expected remaining lifetime increases and R becomes stochastically larger with age. In contrast, light-tailed distributions (e.g., uniform and Gaussian), exhibit expected residual lifetimes that are decreasing functions of age. Finally, for the exponential distribution, age does not affect residual lifetimes

and hence does not provide any useful information for neighbor selection.

Armed with these observations and prior measurement results that demonstrate heavy-tailed user lifetimes in real P2P systems [12], the rest of the chapter explores two simple neighbor-selection methods that rely on age of existing peers to increase network resilience.

4.3. Max-Age Selection

Recall that under uniform selection, each alive user is chosen by peer v with the same probability. To prevent v from connecting to weak neighbors that are about to depart (i.e., users with short remaining lifetimes), this section leverages the heavy-tailed nature of the lifetime distribution $F(x)$ and models the *max-age* neighbor-selection strategy proposed in [12], [41], [77]. In this approach, a joining node v uniformly randomly selects m alive users from the system and chooses the user with the maximal age. It then repeats this procedure k times to obtain its k initial neighbors. The same process is executed every time a dead link is detected.

In what follows in this section, we first analyze the distribution of residuals obtained by the max-age method and then discuss the corresponding isolation probability ϕ.

4.3.1 Residual Lifetime Distribution

Denote by Ω_m the set of m candidate nodes, by U_m the residual lifetime of the max-age user in Ω_m, and by $H^c(x) = P(U_m > x)$ the complementary cumulative distribution function (CCDF) of random variable U_m. Then, we get:

$$H^c(x) = P\Big(R_i > x | A_i = \max_{j \in \Omega_m}\{A_j\}\Big), \tag{89}$$

where A_i is the current age of a user i in Ω_m and R_i is its residual lifetime. Intuitively, (89) states that U_m equals R_i given that user i has the maximum age in Ω_m. Next, following the derivation for the CDF of residual lifetimes under uniform selection in the proof of Theorem 2, the equilibrium age distribution of *existing users* in the system is reduced to

$$F_A(x) = P(A < x) = \frac{1}{E[L]} \int_0^x (1 - F(u))du, \tag{90}$$

where $E[L] < \infty$ as assumed. The following theorem shows that $H^c(x)$ is fully determined by the number of sampled users, lifetime distribution $F(x)$, and age distribution $F_A(x)$.

Theorem 6. *Given that a user's age is larger than that of $m-1$ uniformly selected alive users in the system, its residual lifetime has the following CCDF:*

$$H^c(x) = \frac{m}{E[L]} \int_0^\infty (1 - F(x+y))F_A^{m-1}(y)dy, \tag{91}$$

where $F_A(x)$ is given by (90).

Proof. Recall that A_i represents the maximal user age among m uniformly randomly selected users. It is then clear that the distribution of A_i is:

$$P(A_i < x) = P(\max_{j \in \Omega_m}\{A_j\} < x) = F_A^m(x), \tag{92}$$

where $F_A(x)$ is the equilibrium age distribution of existing users given by (90). Taking the derivative of (294), we immediately get the PDF of A_i:

$$f_{A_i}(x) = \frac{dF_A^m(x)}{dx} = mF_A^{m-1}(x)f_A(x), \tag{93}$$

where $f_A(x) = dF_A(x)/dx$ is the PDF of existing user ages. Assuming an equilibrium

renewal lifetime process, density $f_A(x)$ can be expressed using (90) as:

$$f_A(x) = \frac{dF_A(x)}{dx} = \frac{1 - F(x)}{E[L]}. \tag{94}$$

Substituting (94) into (295), $f_{A_i}(x)$ reduces to:

$$f_{A_i}(x) = \frac{m}{E[L]} F_A(x)^{m-1}(1 - F(x)). \tag{95}$$

Next, conditioning on $A_i = y$, $H^c(x)$ in (89) can be transformed to:

$$H^c(x) = \int_0^\infty P(R_i > x | A_i = y) f_{A_i}(y) dy, \tag{96}$$

where $f_{A_i}(x)$ is given by (296). Observing that $P(R_i > x|A_i = y)$ is equal to $P(L_i - y > x|L_i > y)$ and i could be any user, (96) yields:

$$\begin{aligned} H^c(x) &= \int_0^\infty \frac{P(L_i > x + y)}{P(L_i > y)} f_{A_i}(y) dy \\ &= \int_0^\infty \frac{1 - F(x + y)}{1 - F(y)} f_{A_i}(y) dy, \end{aligned} \tag{97}$$

where $F(x)$ is user lifetime distribution. The last step is to substitute (296) into (297), which then directly leads to (91) after $1 - F(y)$ is canceled. □

Next, we use exponential lifetimes as an example to verify (91). Using $F(x) = F_A(x) = 1 - e^{-\mu x}$, (91) reduces to:

$$H^c(x) = m\mu \int_0^\infty e^{-\mu(x+y)}(1 - e^{-\mu y})^{m-1} dy = e^{-\mu x}. \tag{98}$$

Hence, it follows from (98) that for exponential lifetimes:

$$P(U_m > x) = P(L > x) = e^{-\mu x}, \text{ for any } m \geq 1, \tag{99}$$

which is consistent with the memoryless property of the exponential distribution.

Substituting Pareto lifetimes into (91), we obtain:

$$H^c(x) = \frac{m}{E[L]} \int_0^\infty \Big(1 + \frac{x+y}{\beta}\Big)^{-\alpha} \Big(1 - \Big(1 + \frac{y}{\beta}\Big)^{1-\alpha}\Big)^{m-1} dy, \tag{100}$$

where $E[L] = \beta/(\alpha - 1)$.

Although no closed-form solution for (100) exists in the general case, we next perform a self-check using $m = 1$. Note that for $m = 1$, (100) yields:

$$H^c(x) = \frac{\alpha - 1}{\beta} \int_0^\infty \Big(1 + \frac{x+y}{\beta}\Big)^{-\alpha} dy = \Big(1 + \frac{x}{\beta}\Big)^{1-\alpha}, \tag{101}$$

which indicates that $P(U_1 > x) = P(R > x)$ (i.e., max-age selection with $m = 1$ reduces to single-user uniform selection).

Our next result shows that U_m is stochastically larger than U_{m-1} for any heavy-tailed $F(x)$ and any $m \geq 2$.

Theorem 7. *For any distribution in which larger age implies stochastically larger residuals (i.e., function* (88) *is monotonically increasing in x), the following holds:*

$$P(U_m > x) \geq P(U_{m-1} > x), \quad x \geq 0, m \geq 2. \tag{102}$$

Proof. Denote the maximal user age among m uniformly randomly selected users by:

$$A_m = \max_{j \in \Omega_m} \{A_j\}. \tag{103}$$

It is shown in (294) that the distribution of A_m is given by $P(A_m < x) = F_A^m(x)$. Then, we immediately obtain the following for $m \geq 1$:

$$F_A^{m-1}(x) \geq F_A^m(x) \Leftrightarrow P(A_{m-1} < x) \geq P(A_m < x), \tag{104}$$

which shows that A_m is stochastically larger than A_{m-1}, i.e., $A_m \geq_{st} A_{m-1}$.

Next, denote by:

$$g(y) = P(R > x | A = y), \text{ for fixed } x > 0, \tag{105}$$

the probability that the user residual lifetime is greater than x given that its current age is y. The distribution of U_m can then be transformed from (96) to the following for any fixed $x > 0$:

$$P(U_m > x) = \int_0^\infty g(y) dF_A^m(y) = E[g(A_m)]. \tag{106}$$

Realizing that for any nondecreasing function g, the following holds [91, page 486]:

$$X \geq_{st} Y \Leftrightarrow E[g(X)] \geq E[g(Y)], \tag{107}$$

we easily obtain (102) by using $X = A_m, Y = A_{m-1}$ and substituting (106) into (107). □

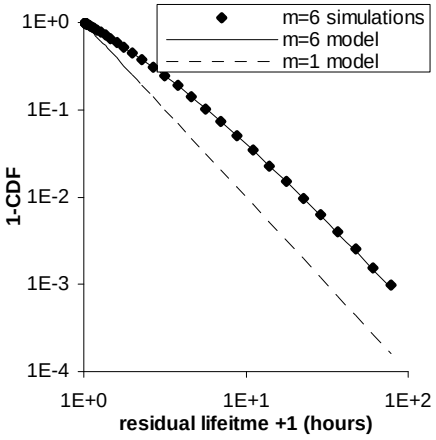
Simulation results in Fig. 10(a) show for $m = 6$ that model (100) is very accurate and random variable U_6 is indeed stochastically larger than R (simulations with other m and those confirming (102) are omitted for brevity). Next, we solve for the expectation of U_m in closed-form for Pareto lifetimes and show the effect of m on the average residual lifetimes of selected neighbors.

Lemma 10. *For Pareto $L \sim 1 - (1 + x/\beta)^{-\alpha}, \alpha > 2$, the expectation of U_m is given by:*

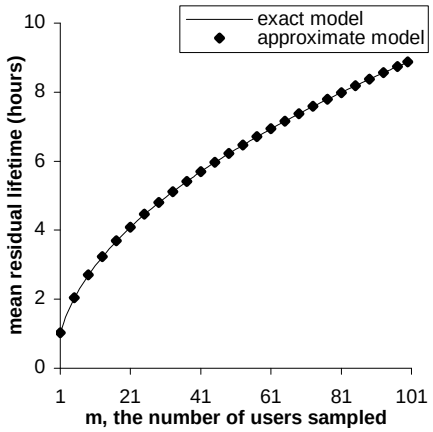
$$E[U_m] = \frac{\beta m! \Gamma(\frac{\alpha-2}{\alpha-1})}{(m(\alpha-1)-1)\Gamma(m - \frac{1}{\alpha-1})}, \quad m \geq 1, \tag{108}$$

where $\Gamma(x)$ is the gamma function. For $\alpha \leq 2$, the expected residual lifetime converges to infinity as system age $\mathcal{T}$ becomes large:

$$\lim_{\mathcal{T} \to \infty} E[U_m] = \infty, \quad m \geq 1. \tag{109}$$



(a) accuracy of (100) with $m = 6$



(b) comparison of (115) to (108)

Fig. 10. Accuracy of models (100) and (115) for Pareto lifetimes with $E[L] = 0.5$ hours and $\alpha = 3$ in a graph with $n = 5,000$ nodes.

Proof. Recall that the expectation of a non-negative random variable U_m can be obtained as:

$$E[U_m] = \int_0^\infty P(U_m > x)dx = \int_0^\infty H^c(x)dx. \tag{110}$$

Substituting $H^c(x)$ from (91) into the above and switching the order of integration variables, we have:

$$E[U_m] = \frac{m}{E[L]} \int_0^\infty \int_0^\infty (1 - F(x+y))dx F_A^{m-1}(y)dy. \tag{111}$$

Using $F(x) = 1 - (1 + x/\beta)^{-\alpha}$ and $F_A(x) = 1 - (1 + x/\beta)^{-\alpha+1}$ and integrating

over x, (111) reduces to:

$$\begin{aligned} E[U_m] &= m\int_0^\infty \big(1+\frac{y}{\beta}\big)^{-\alpha+1}\big(1-(1+\frac{y}{\beta})^{-\alpha+1}\big)^{m-1}dy \\ &= m\beta\int_1^\infty z^{-\alpha+1}\left(1-z^{-\alpha+1}\right)^{m-1}dz \\ &= m\beta\left[{}_2F_1\Big(\frac{1}{1-\alpha},-m;\frac{\alpha-2}{\alpha-1};1\Big)\right. \\ &\left. -\ {}_2F_1\Big(\frac{1}{1-\alpha},1-m;\frac{\alpha-2}{\alpha-1};1\Big)\right], \quad \alpha>2, \end{aligned} \tag{112}$$

where ${}_2F_1(a,b;c;z)$ is the Gauss hypergeometric function [19], which for $z=1$ is:

$${}_2F_1(a,b;c;1) = \frac{\Gamma(c)\Gamma(c-b-a)}{\Gamma(c-a)\Gamma(c-b)}. \tag{113}$$

Using (113) and recalling $\Gamma(m)=(m-1)!$, (112) is transformed into:

$$E[U_m] = m\beta\left(\frac{\Gamma(\frac{\alpha-2}{\alpha-1})m!}{\Gamma(\frac{\alpha-2}{\alpha-1}+m)} - \frac{\Gamma(\frac{\alpha-2}{\alpha-1})(m-1)!}{\Gamma(\frac{\alpha-2}{\alpha-1}+m-1)}\right), \tag{114}$$

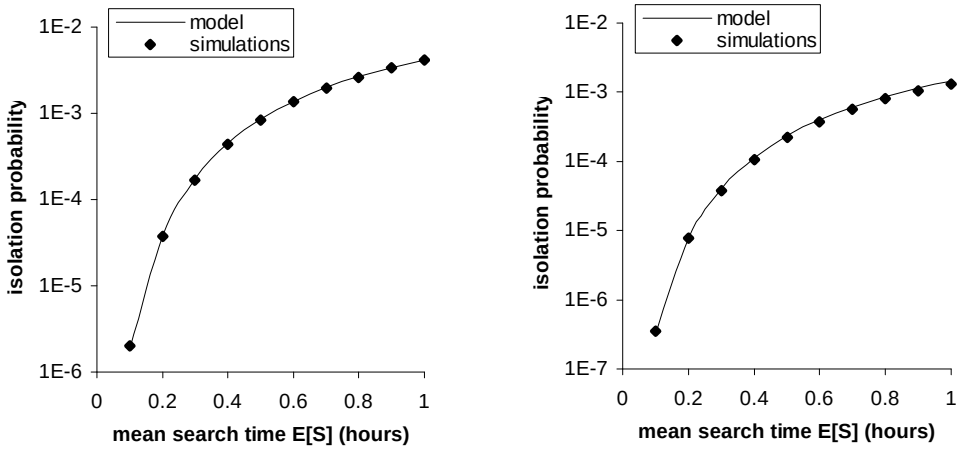
which leads to (108) upon using $\Gamma(x)=(x-1)\Gamma(x-1)$.

For $\alpha \le 2$, recall that $E[U_1]=E[R]=\infty$ under single-user uniform selection. Then it is clear that $E[U_m]=\infty$ for $m\ge 1$ upon invoking Theorem 7. □

To better understand the effect of m on the mean of U_m, we approximate $E[U_m]$ as follows. Setting $c=\Gamma(\frac{\alpha-2}{\alpha-1})$ and expanding the gamma function in the denominator, (108) for $\alpha>2$ yields:

$$E[U_m] \approx cE[L]\Big(m+\frac{1}{\alpha}\Big)^{1/(\alpha-1)}. \tag{115}$$

We next discuss several examples that use (115) with different α. For Pareto lifetimes with $E[L]=0.5$ hours and $\alpha=3$, it can be seen from (115) that $E[U_m]$ follows the curve $0.886(m+0.33)^{0.5} \sim \sqrt{m}$ as $m\to\infty$. However, for smaller α, a more aggressive increase in $E[U_m]$ can be obtained. For $\alpha\to 2$, $E[U_m]\sim m$ is



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

Fig. 11. Comparison of model ϕ to simulations using the max-age selection strategy for Pareto lifetimes with $E[L] = 0.5$ hours and $\alpha = 3$, exponential search times and $k = 7$ in a graph with $5,000$ nodes.

approximately linear, and for $\alpha < 2$, $E[U_m] = \infty$ for any $m \geq 1$ (as before, the last results only holds conditioned on $\mathcal{T} = \infty$). It is also apparent from (115) that as shape parameter α tends to infinity, the impact of m on $E[U_m]$ is weakened and $E[U_m] \to E[L]$, which confirms a well-known fact [42] that Pareto lifetimes with very large α behave as exponential random variables.

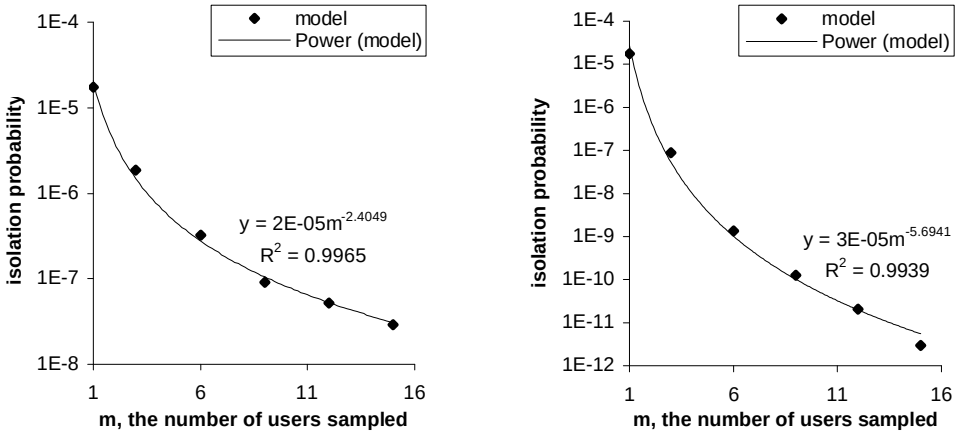
Model (108) is confirmed to be exact using simulations not shown here due to limited space. Fig. 10(b) shows the accuracy of the match between $E[U_m]$ predicted by the exact model (108) and that by the approximate model (115) for $\alpha = 3$. Additional examples with smaller α are omitted for brevity.

4.3.2 Isolation and Resilience

To obtain model ϕ, we approximate the tail of U_m in (91) with its hyper-exponential equivalent in (43) and then compute ϕ by applying Theorem 4 as in Section 4.2.4. Fig. 11 shows ϕ predicted by the model compared to simulations for Pareto lifetimes with

$E[L] = 0.5$ hours, $k = 7$, exponential search delays, and two values of m. As the figure illustrates, the derived result is very accurate and indeed shows inversely proportional dependency between the number of sampled users m and ϕ. The influence of m on isolation probability for Pareto lifetimes is presented more clearly in Fig. 12. As the trendlines show, ϕ is approximately a power-law function m^{-a} for each fixed $E[S]$, where exponent a is $2.4 - 5.7$ in the figure. Thus, for $\alpha = 3$, $m = 10$ sampled users reduce ϕ by a factor of 251 and $m = 30$ by a factor of $3,508$; however, for $\alpha = 2$, $m = 10$ drops ϕ by a factor of $489,000$ and $m = 30$ by a factor of 2.5 billion. Interestingly, while $E[U_m]$ may exhibit an unimpressive growth as a function of m (i.e., linear or slower), the corresponding ϕ demonstrates much faster decay rate and almost always provides significant benefits as m increases.

In systems that do not replace neighbors and $\alpha \to 1$, the limiting isolation probability in (83) is reduced along the corresponding curve in Fig. 12, i.e., proportionally to m^{-a}. Thus, for any finite m, (83) does not qualitatively change its decay rate toward zero as a function of $\gamma = \alpha/(\alpha - 1)$ and leads to no novel discussion. In the next section, however, we develop another neighbor selection framework that guarantees a much stronger result in which ϕ converges to zero for any $1 < \alpha \leq 2$, any number of neighbors $k \geq 1$, and any search delay as system age and size tend to infinity. An additional reason for improving the max-age method in the next section is the difficulty of implementing uniform neighbor selection in decentralized P2P networks without global knowledge at each node. Distributed methods of uniform sampling of users exist [23], [99]; however, they require either k-regular graphs [23] or complex walk patterns [99]. In both cases, max-age selection forces a user to sample m peers to obtain a single neighbor and may not scale well for large m. In contrast, the method we describe below needs only *one* sample per neighbor and operates in graphs with irregular degree distributions.



(a) $\alpha = 3$ (b) $\alpha = 2$

Fig. 12. Influence of m on model ϕ under max-age selection for Pareto lifetimes with $E[L] = 0.5$ hours, exponential search times with $E[S] = 6$ minutes, and $k = 7$.

4.4. Age-Proportional Neighbor Selection

In this section, we first introduce a new neighbor selection strategy that is based on random walks over weighted directed graphs and then deal with the distribution of neighbor residual lifetimes and the corresponding isolation probability.

4.4.1 Random Walks on Weighted Directed Graphs

We start by designing a low-overhead random-walk algorithm whose stationary distribution π ensures that the probability that a user u is selected by another peer is proportional to u's current age. We call the resulting method of choosing neighbors *age-proportional neighbor selection*.

Recall that a directed graph $G = (V, E)$ consists of a vertex set V and edge set E (note that we use notation G instead of $G(t)$ at time t under the assumption that G remains the same while a random walk is performed). Let $u \to v$ represent a directed link $(u, v) \in E$, $N_u^+ = \{v \in V : u \to v\}$ be the set of out-degree neighbors of u, and

$N_u^- = \{v \in V : u \leftarrow v\}$ be the set of in-degree neighbors of u. Further define A_u to be the age of user u and set the weight of each incoming edge $v \rightarrow u$ at node u to be u's age normalized by the number of in-degree neighbors:

$$w(v,u) = \frac{A_u}{|N_u^-|}. \tag{116}$$

It then follows that the in-degree d_u^- of u is simply its age:

$$d_u^- = \sum_{v \in N_u^-} w(v,u) = A_u, \tag{117}$$

and its out-degree d_u^+ is the sum of normalized ages of its out-degree neighbors:

$$d_u^+ = \sum_{v \in N_u^+} w(u,v) = \sum_{v \in N_u^+} \frac{A_v}{|N_v^-|}. \tag{118}$$

Then, age-proportional random walks are executed by alternating between walking along incoming and outgoing edges as we describe next. Given that the walk is currently at node u, the first jump is performed to an *in-degree* neighbor h of u, $h \in N_u^-$, with probability

$$p_{uh} = \frac{w(h,u)}{d_u^-}. \tag{119}$$

The second jump is performed to an *out-degree* neighbor v of h with probability:

$$p_{hv} = \frac{w(h,v)}{d_h^+}. \tag{120}$$

It is clear that the transition probability from u to v is $p_{uv} = \sum_{h \in N_u^-} p_{uh} p_{hv}$. After the two jumps, v becomes the current node and this procedure repeats. Each step consists of two jumps, the node reached after l steps is selected as the neighbor of the current user. As shown in [100], the stationary distribution of this random walk is given by $\pi = (\pi_u)$, where $\pi_u = d_u^- / \sum_{v \in V} d_v^-$. Recalling (117), we immediately obtain

that age-proportional random walks achieve the desired distribution:

$$\pi_u = \frac{A_u}{\sum_{v \in V} A_v}, \text{ for all } u \in V. \tag{121}$$

The starting point of a random walk is determined as follows. Each new user executes a random walk starting from an alive user obtained through bootstrap, while each existing user uniformly randomly selects one of its currently alive out-degree neighbors as the initial point of the walk. Note that if a node does not have any incoming edges, it will never be selected by our walk. To avoid this situation, we alternate between ending walks with an in-degree and an out-degree jump, which gives new users an opportunity to receive incoming edges. Generally speaking, the walk needs to be longer than the mixing time of the chain corresponding to the underlying graph [53]. Simulations below use random walks of $l = 10$ steps as further increasing l does not result in measurable improvements in π for the cases considered in this chapter

4.4.2 Residual Lifetime Distribution

Denote by Z the residual lifetimes of neighbors obtained by age-proportional neighbor selection and by $H^c(x) = P(Z > x)$ its CCDF. We then obtain the distribution of Z in the next theorem.

Theorem 8. *Given that mean $E[L] < \infty$ and variance $Var[L] < \infty$, neighbor residual lifetime Z has the following CCDF:*

$$H^c(x) = \frac{1}{E[L]E[A]} \int_0^\infty y(1 - F(x + y))dy, \tag{122}$$

where $E[A]$ is the mean age of an alive user.

Proof. Denote by A_i the age of node i, $i \in V$, where V is the set of alive users, and

by A_s the age of the user sampled by age-proportional selection. Further denote by $f_{A_s}(x)$ the PDF of A_s such that for infinitely small dx:

$$f_{A_s}(x)dx = P(x < A_s < x + dx). \tag{123}$$

Conditioning on ages A_i for all $i \in V$, (123) is transformed into the following under age-proportional selection:

$$f_{A_s}(x)dx = \frac{x \sum_{i \in V} \mathbf{1}_{x<A_i<x+dx}}{\sum_{i \in V} A_i}, \tag{124}$$

where $\mathbf{1}_X$ is an indicator function such that $\mathbf{1}_X = 1$ if X is true and $1_X = 0$ otherwise. In a system with a large number of users, we can then invoke the law of large numbers to obtain:

$$f_{A_s}(x)dx = \frac{x|V|f_A(x)dx}{|V|E[A]}, \tag{125}$$

where $E[A]$ is the mean age of an alive user, $f_A(x)$ is its PDF given by (94), and $|V|$ is the number of nodes in set V. It immediately follows that:

$$f_{A_s}(x) = \frac{xf_A(x)}{E[A]}, \tag{126}$$

which shows that the age distribution of sampled users is actually the spread distribution [91] of A, i.e., a convolution of two equilibrium age distributions $f_A(x)$ given in (94). This means that $A_s = A + A$, which implies that Z is the residual of a renewal process whose cycle lengths are given by random variable A.

Next, following the derivation in (297) and using (126), we obtain the CCDF of

Z as:

$$H^c(x) = P(Z > x) = \int_0^\infty P(Z > x | A_s = y) f_{A_s}(y) dy$$
$$= \int_0^\infty \frac{1 - F(x+y)}{1 - F(y)} \frac{y}{E[A]} f_A(y) dy, \tag{127}$$

which leads to (122) upon substituting (94) into (127) and then removing the common divisor $1 - F(y)$. □

It is easy to show that for exponential lifetimes, (122) reduces to $1 - F(x)$, again confirming the memoryless property of exponential distributions. For Pareto lifetimes, the CCDF of Z is also very simple given our informal discussion in the previous proof. Since Z is the residual of a renewal process with Pareto cycle length A, we obtain that Z is also Pareto with shape that is smaller than that of A by 1. Since A's shape parameter is $\alpha - 1$, Z exhibits shape $\alpha - 2$. We formally prove this in the next lemma.

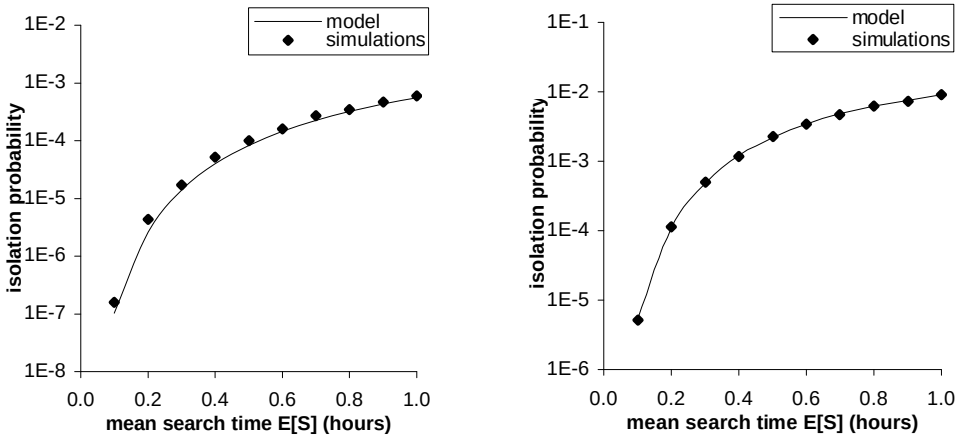
Lemma 11. *For Pareto lifetimes $L \sim 1 - (1 + x/\beta)^{-\alpha}$ with $\alpha > 2$, the CCDF of Z is given by:*

$$H^c(x) = \left(1 + \frac{x}{\beta}\right)^{-(\alpha-2)}. \tag{128}$$

For $1 < \alpha \leq 2$, Z converges in probability to ∞ as system age $\mathcal{T}$ and size n both tend to ∞. For $\alpha > 3$, the expectation of Z is $E[Z] = \beta/(\alpha - 3)$ and for $1 < \alpha \leq 3$ it is $E[Z] = \infty$.

Proof. For Pareto lifetimes, straightforward integration of (122) leads to:

$$H^c(x) = \frac{1}{E[L]E[A]} \int_0^\infty y \left(1 + \frac{x+y}{\beta}\right)^{-\alpha} dy$$
$$= \frac{\beta^2}{E[L]E[A]} \frac{(1 + \frac{x}{\beta})^{-\alpha+2}}{(\alpha-2)(\alpha-1)}, \quad \alpha > 2, \tag{129}$$



(a) $\alpha = 3$ (b) $\alpha = 5$

Fig. 13. Comparison of model ϕ to simulations under age-proportional random walks for Pareto lifetimes, $E[L] = 0.5$ hours, $\beta = (\alpha - 1)E[L]$, exponential search delays, and $k = 7$ in a graph with $n = 8,000$ nodes.

which gives us the desired result by recalling that $E[L] = \beta/(\alpha - 1)$ and $E[A] = \beta/(\alpha - 2)$. For $1 < \alpha \leq 2$, we have $E[A] = \infty$. In this case, it is known from [18] that residuals Z converge in probability to ∞ as system $\mathcal{T}$ and size n become large. Note that $\mathcal{T} \to \infty$ is needed to obtain the limiting distribution (94) of age A with $E[A] = \infty$ and $n \to \infty$ is needed for age A_i of selected user i to become the spread of A during the process of selecting neighbors from the current user population.

For $\alpha > 2$, integrating (128) leads to:

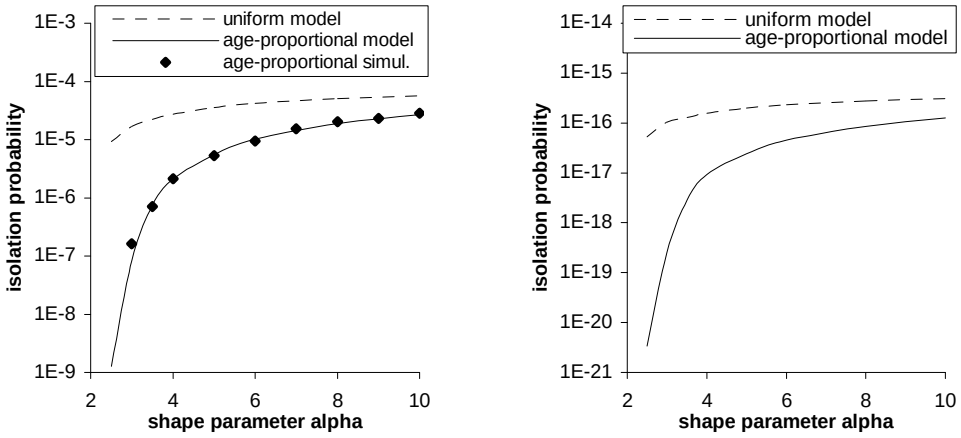
$$E[Z] = \int_0^\infty H^c(x)dx = \int_0^\infty \left(1 + \frac{x}{\beta}\right)^{-\alpha+2} dx$$
$$= \begin{cases} \dfrac{\beta}{\alpha - 3} & \alpha > 3 \\ \infty & 2 < \alpha \leq 3 \end{cases}. \tag{130}$$

For $1 < \alpha \leq 2$, it is easy to obtain that $E[Z] = \infty$ since Z converges in probability to ∞. □

Note that for $\alpha > 2$, the PDF of Z is completely monotone and thus suitable for our hyper-exponential model. Also notice that Z is stochastically larger than residual lifetimes R under uniform selection for all choices of α. In fact, Z shifts the shape of the Pareto distribution from α to $\alpha - 2$, which is not achievable under max-age selection even as $m \to \infty$. Furthermore, for $1 < \alpha \leq 2$, residuals Z tend to a defective random variable with all mass concentrated at $+\infty$ as system size and age become infinite. This shows that in asymptotically large systems, Z exceeds any lifetime L with probability 1 and no user suffers isolation (more on this below).

4.4.3 Isolation and Resilience

To obtain model ϕ under age-proportional neighbor selection, we fit the distribution of Z shown in (128) with its hyper-exponential equivalent and then invoke Theorem 4 to solve for ϕ. Next, we test the accuracy of model ϕ in simulations where $n = 8,000$ nodes join and leave the system at random instances and each node performs age-proportional random walks to find its neighbors. As shown in Fig. 13, simulation results are very close to the values predicted by theoretical ϕ. Examples showing the relationship between of ϕ and α are presented in Fig. 14. As shown in Fig. 14(a), simulation results are consistent with model ϕ under a variety of values α that allow quick simulations and do not require very large $\mathcal{T}$ or n (i.e., $\alpha \geq 3$). It is interesting to observe in the figure that as α decreases, the gap between ϕ under age-proportional random walks and that under uniform selection drastically increases and reaches a factor of 10^4 for $\alpha = 2.5$. This shows that age-proportional random walks are extremely effective in systems with very heavy-tailed lifetimes (i.e., α below 2.5). Fig. 13(b) shows that the same conclusion holds for $E[S] = 3.6$ seconds, in which case ϕ is on the order of 10^{-20} and only allows computation using the model since simulations are impractical for such small probabilities.



(a) $E[S] = 6$ minutes (b) $E[S] = 3.6$ seconds

Fig. 14. Impact of α on ϕ under uniform selection and under age-proportional random walks for Pareto lifetimes, $E[L] = 0.5$ hours, $\beta = (\alpha - 1)E[L]$, exponential search delays, and $k = 7$.

The most intriguing result shown in Fig. 14 is that ϕ tends to 0 as α converges to 2 from above. However, as before, this convergence requires that system age tend to infinity. In addition, the following result states that system size n must also be infinite to obtain $\phi = 0$.

Theorem 9. *For an equilibrium system, Pareto lifetimes with $\alpha > 2$, and infinitely large search delay (i.e., $S = \infty$), isolation probability ϕ under age-proportional neighbor selection is given by:*

$$\phi = \frac{k!}{(\theta + 1) \times \ldots \times (\theta + k)}, \tag{131}$$

where $\theta = \alpha/(\alpha - 2)$. For $\alpha \to 2$ and fixed k, (131) converges to 0 as $\Theta(\theta^{-k})$.

For Pareto lifetimes with $1 < \alpha \leq 2$, any number of neighbors $k \geq 1$, and any type of search delay (including $S = \infty$), the isolation probability under age-proportional neighbor selection converges to zero as system age $\mathcal{T}$ and size n approach infinity: $\lim_{n\to\infty} \lim_{\mathcal{T}\to\infty} \phi = 0$.

Proof. Let us consider ϕ for $\alpha > 2$ and $S = \infty$ first. Recall that if $S = \infty$, the first hitting time T is the maximum residual lifetime among k neighbors. Using (128), we then readily get the following for $\alpha > 2$:

$$P(T < x) = P(Z < x)^k = \left[1 - \left(1 + \frac{x}{\beta}\right)^{-\alpha+2}\right]^k. \tag{132}$$

Following derivations in the proof of Theorem 5, it is easy to obtain:

$$\begin{aligned}\phi &= \int_0^\infty P(T < x) f(x) dx = \frac{k!\Gamma(1+\theta)}{\Gamma(1+k+\theta)} \\ &= \frac{k!}{(\theta+1) \times \ldots \times (\theta+k)},\end{aligned} \tag{133}$$

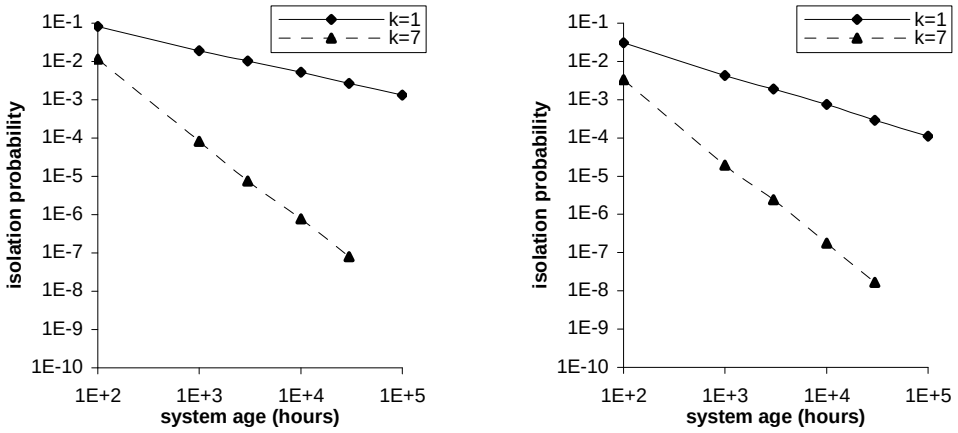
where $f(x) = \alpha(1 + x/\beta)^{-\alpha-1}/\beta$ is the PDF of Pareto L, $\theta = \alpha/(\alpha - 2)$, and $\Gamma(x)$ is the gamma function.

As $\alpha \to 2$, it is clear from (133) that $\theta \to \infty$, which makes ϕ approach 0 as $\Theta(\theta^{-k})$ for fixed k.

For $1 < \alpha \leq 2$, it has been shown in Lemma 11 that $P(Z < x) \to 0$ for any $x > 0$ as system age $\mathcal{T}$ and system size n approach infinity. Supposing $k = 1$, we readily obtain $\phi = P(Z < L) \to 0$. Noticing that ϕ for any $k \geq 2$ (including $S = \infty$ and $S < \infty$) is smaller than that for $k = 1$, we immediately establish Theorem 9. □

Note that Theorem 9 is a much stronger result than Theorem 5 since ϕ under uniform selection does *not* asymptotically approach 0 for any fixed $\alpha \in (1, 2]$. However, the asymptotic result of this section is more difficult to achieve since it requires not only an equilibrium system, but also an infinitely large user population.

We finish this section by examining age-proportional random walks under finite $\mathcal{T}$ and n using several values of $1 < \alpha \leq 2$. For such cases, recall from Lemma 11 that Z converges in probability to ∞; however, initial analysis shows that the convergence rate of $Z \to \infty$ and $\phi \to 0$ can only be expressed using complex Appell



(a) $\alpha = 1.5, S = \infty$ (b) $\alpha = 1.2, S = \infty$

Fig. 15. Simulation results of ϕ under age-proportional selection as system age $\mathcal{T}$ and size n increase for Pareto lifetimes with $E[L] = 0.5$ hours.

hypergeometric functions [19] of $\mathcal{T}$ and n for which no closed-form expansion exists. We leave this task for future work and instead show simulations of ϕ in Fig. 15 as $\mathcal{T}$ becomes large (n is kept equal to $\mathcal{T}/10$). For both values of α, the figure shows that ϕ monotonically decreases as system age $\mathcal{T}$ increases. In fact, for $k = 7$, the system achieves isolation probability below 10^{-7} without replacing neighbors at $\mathcal{T} = 30,000$ hours and $n = 3,000$ users. Additional simulations with $k = 7$ suggest that increasing n to over 1 million users and keeping the age around 1 year will produce ϕ sufficiently small for most large-scale networks today.

4.5. Summary

This chapter derived a general model of resilience for unstructured P2P networks under heavy-tailed user lifetimes and formally analyzed two age-dependent neighbor-selection techniques. Our results show that the proposed random-walk method may achieve *any* desired level of resilience without replacing the neighbors as long as Pareto shape parameter $1 < \alpha \leq 2$ and system size n and age $\mathcal{T}$ are sufficiently

large. This indicates that P2P systems under proposed neighbor selection and very heavy-tailed lifetimes (i.e., $\alpha < 2$) become progressively more resilient over time and asymptotically tend to an "ideal" system that never disconnects as users join the network.

CHAPTER V

NODE IN-DEGREE AND JOINT IN/OUT-DEGREE

5.1. Introduction

Chapter IV focused on the *out-degree* of each user and did not consider the increased resilience arising from additional *in-degree* edges arriving in the background to each user during its stay in the system.

In this chapter, we overcome this shortcoming and build a complete closed-form model characterizing the evolution of in-degree in unstructured systems under the assumption of uniform neighbor selection. We first show that under certain mild assumptions, the edge arrival process to each user tends to a Poisson distribution when system size becomes sufficiently large, which is consistent with recent observation of this phenomenon in certain real networks [81]. We then derive the transient distribution of in-degree as a simple function of $F(x)$, *including cases with non-exponential peer lifetimes*, and show that users who stay online longer quickly accumulate non-trivial in-degree and become much more resilient to isolation over time. This outcome was intuitively expected as it makes sense that current unstructured P2P networks have been designed such that users with more contribution to the system (i.e., longer lives) become better connected over time and provide more search capabilities to their neighbors. In contrast, the original model of [43] showed that P2P users became progressively more susceptible to isolation as their age increased.

We finish the chapter by combining the in and out-degree isolation models into a single approximation that clearly shows the contribution of in-degree to the resilience of the graph. Denoting by ϕ the isolation probability of a user (i.e., loss of all

neighbors within its lifetime) and by ϕ_{out} the same metric with only the out-degree being considered [43], we show that for exponential $F(x)$ the following holds as search delays become asymptotically small (i.e., tend to zero):

$$\phi = \frac{1 - e^{-2k}}{2k}\phi_{out}, \tag{134}$$

where k is the initial number of neighbors obtained by each arriving user. This result illustrates that the amount of improvement from the in-degree amounts to approximately a factor of $2k$ reduction in the isolation probability. We also observe from our closed-form Markov-chain model that for non-negligible search delays, ratio ϕ_{out}/ϕ is often much larger than implied by (134), which suggests that (134) may be a worst-case upper bound on ϕ. We finish the chapter with examples that demonstrate this effect.

5.2. Edge Arrival

Before analyzing node in-degree under uniform selection, we study the process of edge arrival into each user since this determines both the rate at which the user accumulates incoming neighbors and the stationary in-degree distribution. Our neighbor churn model prescribes that each joining user find k random out-degree neighbors and then continuously replace them as they fail, as in [43]. Define *initial edges* to be those added when users arrive in the system and *replacement edges* to be those added in response to neighbor failures.

Assumption 3. *The number of neighbors k a user selects upon joining the system is a constant for all n.*

This assumption often holds in unstructured P2P networks where individual users are unaware of system size (e.g., Gnutella) and some structured P2P networks

with constant node degree (e.g., de Bruijn [51]).

5.2.1 Definitions

Considering the time-limiting behavior of the system (i.e., $t \to \infty$), the rest of the chapter assumes that user ON/OFF process $Z_i := \{Z_i(t)\}_{t\geq 0}$ (see Chapter III) are *stationary* alternating renewal processes on time interval $[0, \infty)$, for $i = 1, 2, \ldots, n$. Denote by $\{\tau_{i,m}\}_{m=1}^{\infty}$ arrival times of user i. Then, $\tau_{i,m+1} = \tau_{i,m} + L_{i,m} + D_{i,m}$, for $m \geq 1$. To ensure stationarity, let $\tau_{i,1} := L_i^e + D_i$ w.p. a_i and otherwise $\tau_{i,1} := D_i^e$, where L_i^e has the equilibrium distribution of $F_i(x)$ and D_i^e has that of $G_i(x)$. Define $M_i(t)$ to be the number of arrivals of user i in interval $[0, t]$:

$$M_i(t) := \sum_{m=1}^{\infty} \mathbf{1}_{\tau_{i,m}\in[0,t]}, \tag{135}$$

whose expectation (due to stationarity) is $E[M_i(t)] = \lambda_i t$ for any $t \geq 0$, where λ_i is given in (3).

Recall that in our resilience model, each user i has k out-degree neighbors, which are either dead (i.e., a replacement is being sought) or alive at any given time. Let $Y_i^c := \{Y_i^c(t)\}$ be an alternating process indicating the state of i's link c:

$$Y_i^c(t) := \begin{cases} 1 & \text{out-link } c \text{ of user } i \text{ is ALIVE at } t \\ 0 & \text{otherwise (DEAD)} \end{cases}, \tag{136}$$

for $c = 1, \ldots, k$. If i is offline at t, all of its links are considered dead. The out-degree of user i at time t is simply $\sum_{c=1}^{k} Y_i^c(t)$. Whenever Y_i^c transitions from DEAD to ALIVE, user i delivers one edge into the system (i.e., performs one selection). Thus, processes $\{Y_i^c\}_{i,c}$ determine the edge-generation rate of individual users.

As illustrated in Fig. 16, link c becomes ALIVE at arrival times $\{\tau_{i,m}\}_{m\geq 1}$ and then alternates between DEAD/ALIVE states during i's ON periods. Note that

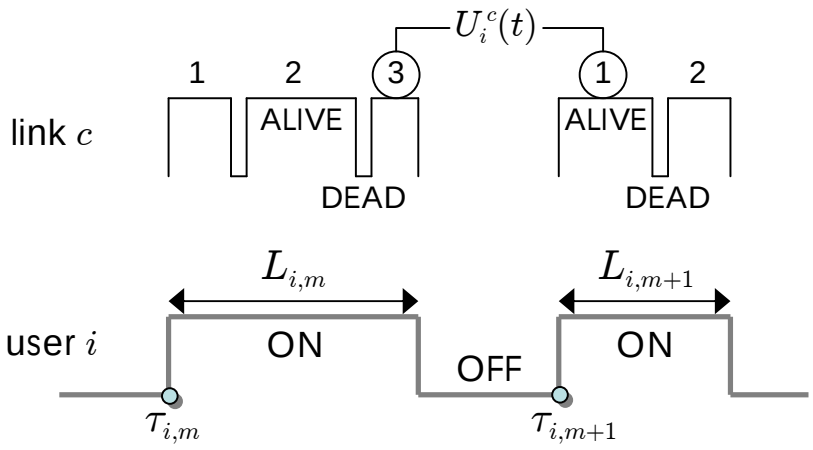


Fig. 16. Process $\{Y_i^c(t)\}$ indicates DEAD/ALIVE behavior of the c-th out-link of user i. Process $\{U_i^c(t)\}$ counts the number of DEAD→ALIVE transitions within the current ON cycle of i.

ALIVE durations of Y_i^c are neighbor residual lifetimes and DEAD durations are search delays for finding replacement neighbors, with the exception of the very last ALIVE cycle in each ON period, which is terminated by i's departure rather than neighbor failure. To save space, we assume that search delays (they can be accommodated by changing link residual lifetime $R(t)$ to $R(t) + S(t)$, where $S(t)$ is the search delay at t) are negligible compared to $L_{i,m}$ and do not explicitly model their effect on the in-degree process.

The figure also shows right-continuous process $\{U_i^c(t)\}$, which is the number of transitions DEAD→ALIVE within the current ON cycle up to time t. We assume $U_i^c(\tau_{i,m}) = 1$ for all $m \geq 1$, use notation t^- to represent the instant just prior to t, and denote by

$$U_i^c(\tau_{i,m+1}^-) = \sup_{\tau_{i,m} \leq t < \tau_{i,m+1}} U_i^c(t) \tag{137}$$

the number of selections for link c in the m-th ON cycle. It then follows that $U_i^c := \{U_i^c(t)\}$, for all c and i, are stationary processes since they are functions of stationary

$\{Z_i(t)\}$.

Note that $U_i^c := \{U_i^c(t)\}_{t\geq 0}$, for all c and i, are stationary processes since they are functions of stationary processes Z_i. We assume that the initial distribution of U_i^c at time 0 follows its stationary distribution (to this end, image that the 0-th arrival time $\tau_{i,0}$ of user i is placed at random distance to the left of $t = 0$ such that $P(-\tau_{i,0} \leq x)$ is equal to the CDF of $\tau_{i,1}$ and that user i monitors its out-links since $\tau_{i,0} < 0$. With this setup, $U_i^c(0)$ has no jump at time 0 and follows the stationary distribution of U_i^c). Further, observe that if user i starts with an ON period at time 0, $U_i^c(\cdot)$ increases as Y_i^c turns ON in interval $[0, \tau_{i,1})$; otherwise, $U_i^c(\cdot)$ remains the same in that period.

Denote by $\{\delta_{i,z}^c \geq 0\}_{z=1}^{\infty}$ random times at which Y_i^c becomes ALIVE (i.e., an edge is generated by i and delivered to some target peer). Define

$$W_i^c(t) := \sum_{z=1}^{\infty} \mathbf{1}_{\delta_{i,z}^c \in [0,t]} = \sum_{m=1}^{M_i(t)} U_i^c(\tau_{i,m}^-) - U_i^c(0) + U_i^c(t)$$

to be the number of selections for link c in $[0, t]$. Finally, denote by $W_i(t) := \sum_{c=1}^{k} W_i^c(t)$ the number of edges delivered by i into the system in $[0, t]$. Observe that $W(t) = \sum_{i=1}^{n} W_i(t)$ is the number of out-degree edges generated by n users in $[0, t]$, which is the same as the number of in-degree edges received by living users in $[0, t]$.

5.2.2 Edge Creation Process

Our next step is to analyze the rate of edge generation from a given user as $n \to \infty$. Denote by $R(n, \delta_{i,z}^c)$ the residual lifetime of the peer selected by user i at a random instance $\delta_{i,z}^c$. Invoking Theorem 2, we next examine $R(n, \delta_{i,z}^c)$.

Lemma 12. *Given Assumption 2 and uniform selection, fix user i and $t > 0$. Then,*

(1) Random variables $\{U_i^c(\tau_{i,m}^-)\}$ and $\{W_i(t)\}$ are uniformly integrable in n;

(2) For arbitrary $t^\star > 0$, residuals $\{R(n, \delta^c_{i,z})\}_{z \le W^c_i(t), t \le t^\star}$ of selected neighbors converge in distribution as $n \to \infty$ to i.i.d. r.v.'s with CDF $H(x)$ in (35);

(3) The average number of selections per ON cycle for each out-link c converges as $n \to \infty$ to

$$E[U^c_i(\tau^-_{i,m})] \to 1 + \sum_{r=1}^{\infty} \int_0^{\infty} H^{*r}(x) dF_i(x) < \infty, \tag{138}$$

for $m \ge 2$, where $H^{*r}(x)$ is the r-th convolution of $H(x)$ and $F_i(x)$ is the lifetime distribution of i.

Proof. We prove each of the statements in sequence.

5.2.2.1 Uniform Integrability

Part (1) paves the way to establish convergence results on moments of associated variables. The key aspect of this proof is to show that $U^c_i(t)$ is stochastically smaller than some variable $\widehat{U}^c_i(t)$, which is independent of n. This independence automatically implies uniform integrability of both $\widehat{U}^c_i(t)$ and all r.v. stochastically smaller than it. The major impediment to achieving this is that uniform selection allows user i to repeated connections to the same user, which creates dependency of residuals of acquired neighbors. While for $n \to \infty$ this dependency diminishes, our analysis in this proof takes it into account and creates a foundation that will be used in the derivations that follow in the next section.

To proceed, call a *new* neighbor of user i if it is different from any previous selections that i makes for *all* of k out-links since $\tau_{i,m}$. Denote by $H^{(j)}(x) := (l^{(j)})^{-1} \int_0^x (1 - F^{(j)}(u))du$ the residual CDF for user-type j and by $\widehat{H}(x) := \max_{1 \le j \le \mathcal{T}} H^{(j)}(x)$ that is stochastically smaller than *all* distributions $H^{(j)}(x)$ for $j = 1, \ldots, \mathcal{T}$. We now create a virtual process for node i whose number of neighbor selections by time t within the

current ON period is $\widehat{U}_i^c(t) \geq^{st} U_i(t)$. We achieve this by letting i acquire *new* selections with residuals drawn from $\widehat{H}(x)$ and *old* (as opposed to new) selections with residuals deterministically set to 0. Indeed, this represents the worst-case scenario for all neighbor choices.

Now, define η_z^c, $z \geq 1$, to be random times at which user i's out-link c connects to *new* neighbors in the current ON cycle and set $\eta_1^c = \tau_{i,m}$. Denote by $B^c(t)$ the number of new selections for link c in $[\tau_{i,m}, \tau_{i,m} + t]$. Note that $B^c(\eta_z^c) = z$. Further, let Q_z^c count the number of *old* selections in interval (η_{z-1}^c, η_z^c) and set $Q_1^c = 0$. Then, we get

$$\widehat{U}_i^c(t) := B^c(t) + \sum_{z=1}^{B^c(t)} Q_z^c, \tag{139}$$

where Q_z^c has a geometric distribution with success probability p_z^c, i.e., the probability that i gets its z-th new selection for link c, which will be studied next.

Define X_z to be the number of peers that are *alive* for selection at $\eta_z^{c\,-}$ and were chosen as i's *neighbors* in interval $[\tau_{i,m}, \eta_z^c)$ for all of its k links and set $X_1 = 0$. Note that $E\left[X_z | B^c\right] \leq kB^c(\eta_z^{c\,-}) = kz$ for $z \geq 2$. Conditioning on the system size (without i) $N_n^i(\eta_z^c) \geq 1$, the probability p_z^c is then given by

$$p_z^c = 1 - \frac{X_z}{N_n^i(\eta_z^c)}, \tag{140}$$

where $X_z < N_n^i(\eta_z^c)$, and the expectation of Q_z^c is thus

$$\begin{aligned} E\Big[Q_z^c \mid B^c\Big] = E\left[\frac{1-p_z^c}{p_z^c} \mid B^c\right] &= E\Big[\frac{X_z}{N_n^i(\eta_z^c) - X_z} \mid B^c\Big] \\ &\leq E\left[X_z | B^c\right] \leq kz. \end{aligned} \tag{141}$$

This immediately leads to

$$E\Big[\sum_{z=1}^{B^c(t)} Q_z^c\Big] \leq E\Big[\sum_{z=1}^{B^c(t)} kz\Big] \leq kE\left[B^c(t)(B^c(t)+1)\right] < \infty,$$

where $\{B^c(t)\}$ is a renewal process with renewal distribution $\widehat{H}(x)$ independent of n.

We thus get that variables $\{\widehat{U}_i^c(t)\}$ are uniformly integrable in n, which leads to the same conclusion for $\{U_i^c(t)\}$ and $\{U_i^c(\tau_{i,m+1}^-)\}$. This directly implies that $\{W_i(t)\}$ are uniformly integrable in n, where $W_i(t)$ is the number of selections made by i in $[0,t]$.

5.2.2.2 Residuals

We next show that i finds new neighbors w.p. 1 as $n \to \infty$. Since the probability that i selects the same peer at random instances $\delta_{i,z}^c, \delta_{i,z'}^c$ is $1/(N_n^i(\delta_{i,z}^c)N_n^i(\delta_{i,z'}^c))$, the probability b_n that i encounters at least one old user during selections for link c in $[0,t]$ is bounded by

$$b_n \leq E\Big[(n-1)\sum_{z=1}^{W_i^c(t)}\sum_{z'\neq z}^{W_i^c(t)} \frac{1}{N_n^i(\delta_{i,z}^c)N_n^i(\delta_{i,z'}^c)} | N_n^i \geq 1\Big].$$

Given a stationary system, the above yields

$$\lim_{n\to\infty} b_n = \lim_{n\to\infty} E\Big[\frac{n-1}{N_n^i(t)^2} | N_n^i \geq 1\Big] E\left[W_i^c(t)^2\right] = 0, \tag{142}$$

where $E[1/(N_n^i(t))^2 | N_n^i \geq 1] = \Theta(\mu_n^{-2}) = \Theta(n^{-2})$ from Lemma 4 and $E[W_i^c(t)^2] < \infty$. It follows almost surely that all neighbors selected by user i for its k links in $[0,t]$ are new as $n \to \infty$. This immediately leads to the fact that residual lifetimes $R(n, \delta_{i,z}^c)$ at random $\delta_{i,z}^c \leq t$ are independent (as different users are independent of each other) and have the same limiting distribution of residual $R(n,t)$ selected at fixed t (due to stationarity of Z_i), where $\lim_{n\to\infty} P(R(n,t) \leq x) = H(x)$ is given in (35).

5.2.2.3 Edges

The rest of the proof directly follows from renewal theory. Denote by $\{B(t)\}_{t\geq 0}$ a pure renewal process with waiting times $R_r \sim H(x)$ for $r \geq 1$. We then have

$$E[B(t)] = \sum_{r=0}^{\infty} P(B(t) > r) = 1 + \sum_{r=1}^{\infty} H^{*r}(t). \tag{143}$$

Noting that $L_{i,m} \sim F_i(x)$ is independent of $\{B(t)\}$, the mean number of cycles before user departure is given by

$$\lim_{n\to\infty} E[U_i^c(\tau_{i,m}^-)] = E[B(L_{i,m})] = \int_0^{\infty} E[B(t)]dF_i(t),$$

which establishes (138). □

It is now clear that $\{U_i^c(t)\}_{t\geq 0}$ converge in distribution as $n \to \infty$ to a *stationary regenerative* processes with regeneration epochs $0 < \tau_{i,1} \leq \tau_{i,2} \ldots$ Recalling that $W_i(t)$ is uniformly integrable and that $E[W_i(t)] = kE[W_i^c(t)]$, the next result is directly obtained from Smith's theorem for stationary regenerative processes.

Lemma 13. *With Assumptions 2-3, uniform selection, and fixed user i and time t, the expected number of edges from i in $[0, t]$ converges*

$$\lim_{n\to\infty} E[W_i(t)] = \lambda_i t(k + \theta_i) < \infty, \tag{144}$$

where λ_i is in (3) and $\theta_i := k\sum_{r=1}^{\infty}\int_0^{\infty} H^{*r}(x)dF_i(x)$ is the mean number of replacement edges created per session of i.

This result can be interpreted as each user generating $k + \theta$ edges per arrival interval $l_i + d_i$ and segment $[0, t]$ containing $t\lambda_i$ such intervals on average. We leverage this observation in the next subsection.

5.2.3 Edge Arrival Process

Now, given a set of n participating users, our approach is to set aside a certain peer of interest and examine edge-arrivals to this peer during its lifetime from n other users under uniform selection. Without loss of generality, we study edge-arrivals from users $1, \ldots, n$ to special user 0 conditioned on its being alive during all manipulations. Indeed, since ON/OFF periods of Z_0 are independent of each other and the edge-arrival process is independent of Z_0, this analysis directly generalizes to other users.

Define $I_{i,z}^c$ to be a Bernoulli r.v. indicating whether user 0 is chosen by $i \geq 1$ at time $\delta_{i,z}^c$, where $c = 1, \ldots, k$ and $z \geq 1$. Conditional on $N_n := \{N(n,t)\}$ and Y_i^c, the probability that $I_{i,z}^c = 1$ under uniform selection is

$$p_{i,z}^c := P(I_{i,z}^c = 1 | N_n, Y_i^c) = \frac{1}{N_n^i(\delta_{i,z}^c) + 1}, \tag{145}$$

where $N_n^i(t) := \sum_{j=1, j\neq i}^n Z_j(t)$ is the population size excluding user i (to avoid self-loops) and not counting the always-alive user 0, which is explicitly added in the denominator of (145). Note that $I_{i,z}^c$ are *conditionally independent* given N_n and all processes $\{Y_i^c\}_{i,c}$, i.e.,

$$P\Big(\bigcap_{i,z,c} [I_{i,z}^c = 1] | N_n, \{Y_i^c\}_{i,c}\Big) = \prod_{i,z,c} p_{i,z}^c$$

for $1 \leq i \leq n, z \geq 1, 1 \leq c \leq k$. Then, the number of edges delivered by user i to user 0 in interval $[0,t]$ is $\xi_{ni}(t) := \sum_{c=1}^k \sum_{z:\delta_{i,z}^c \leq t} I_{i,z}^c$. Finally, the number of edges from the system to user 0 in $[0,t]$ is

$$\xi_n(t) := \sum_{i=1}^n \xi_{ni}(t). \tag{146}$$

The properties of process $\xi_n := \{\xi_n(t)\}$ are given next.

Theorem 10. *Under Assumptions 2-3 and uniform selection, the point process* ξ_n

defined in (146) *converges in distribution as* $n \to \infty$ *to a Poisson process* ξ *with constant rate:*

$$\nu := \frac{k+\theta}{l}, \tag{147}$$

$\theta := k \sum_{r=1}^{\infty} \int_0^{\infty} H^{*r}(x) dF(x)$ *is the mean number of replacement edges generated per user ON cycle and is independent of* n, $F(x)$ *is the lifetime CDF shown in* (33)*, and* l *is the mean lifetime in* (34)*.*

Proof. We set ξ to be a Poisson process with finite rate ν. It has been shown in [69, Proposition 3.22] that ξ_n converges in distribution to ξ under the following constraints:

(1) $\forall t > 0 : \lim_{n\to\infty} E[\xi_n(t)] = E[\xi(t)] = \nu t$; and

(2) $\forall t > 0 : \lim_{n\to\infty} P(\xi_n(t) = 0) = P(\xi(t) = 0) = e^{-\nu t}$.

We set ξ to be a Poisson process with rate ν and establish these conditions next.

5.2.3.1 Continuity

This condition is trivially met since the first, and thus the remaining, arrival times of any user i have an absolutely continuous distribution, which is ensured by stationarity and non-lattice lifetime distributions.

5.2.3.2 Mean Convergence

Our next step is to show that $\lim_{n\to\infty} E[\xi_n(t)] = \nu t < \infty$. Write

$$\begin{aligned} E[\xi_n(t)|N_n, \{Y_i^c\}_{i,c}] &= E\Big[\sum_{i=1}^{n}\sum_{c=1}^{k}\sum_{z:\delta_{i,z}^c \le t} I_{i,z}^c\Big] \\ &= \sum_{i=1}^{n}\sum_{c=1}^{k}\sum_{z:\delta_{i,z}^c \le t} p_{i,z}^c. \end{aligned} \tag{148}$$

Leveraging Lemma 4 for uniform integrability of $n/N(n)$, stationarity of users, and Lemma 13 for the convergence of $E[W_i(t)]$, we have after unconditioning of (148)

$$\begin{aligned}\lim_{n\to\infty} E[\xi_n(t)] &= \lim_{t\to\infty} E\Big[\sum_{i=1}^{n}\sum_{c=1}^{k}\sum_{z:\delta_{i,z}^c\le t} p_{i,z}^c\Big] \\ &= \lim_{n\to\infty} E\Big[\sum_{i=1}^{n} W_i(t)\Big] E[p_{1,1}^1] \\ &= \lim_{n\to\infty} \frac{\sum_{i=1}^{n}(k+\theta_i)\lambda_i t}{\sum_{i=1}^{n} a_i}, \end{aligned} \tag{149}$$

where θ_i is given in (144). By Lemma 5, the above reduces to

$$\lim_{n\to\infty} E[\xi_n(t)] = \frac{t}{l}\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\lambda_i(k+\theta_i)}{\sum_{i=1}^{n}\lambda_i} = \frac{t}{l}(k+\theta), \tag{150}$$

where the last limit holds a.s. Since θ is independent of n, we know that (150) is finite.

5.2.3.3 Probability Convergence

In this step, we show that $P(\xi_n(t)=0)\to e^{-\nu t}$ as $n\to\infty$. Since $I_{i,z}^c$ are conditionally independent given N_n and Y_i^c, we have

$$P(\xi_n(t)=0|N_n,\{Y_i^c\}_{i,c}) = e^{-B_n}, \tag{151}$$

where

$$\begin{aligned} B_n &= -\sum_{i=1}^{n}\sum_{c=1}^{k}\sum_{z:\delta_{i,z}^c\le t}\log(1-p_{i,z}^c) \\ &= -\frac{1}{E[N(n)]}\sum_{i=1}^{n}\sum_{j=1}^{W_i(t)} E[N(n)]\log(1-p_{ij}), \end{aligned} \tag{152}$$

where p_{ij} is the probability for user i to choose v during its j-th selection in $[0,t]$, similar to one shown in (145). Note that $P(\xi_n(t)=0)$ is then simply $E[e^{-B_n}]$. We

next show that $B_n \to \nu t$ in probability and that B_n is uniformly integrable, from which the desired result follows immediately.

Define

$$f(x) = -E[N(n)] \log\Big(1 - \frac{1}{x+1}\Big) \tag{153}$$

to be a continuous, monotonically decreasing function of x for $x \geq 1$. We next sketch a proof for $f(N(n)) \to 1$ in r-th mean for all $r \geq 1$. Following Lemma 4, $E[f(N(n))]$ can be split into two expectations conditioned on $\mathcal{A} = |N(n)/E[N(n)] - 1| \leq 1 + \delta$ and $\mathcal{B} = |N(n)/E[N(n)] - 1| > 1 + \delta$. For the first condition $\mathcal{A}$, which holds with w.p. $1 - o(1)$ as $n \to \infty$, it is easy to show that the corresponding term $E[f(N(n))|\mathcal{A}]P(\mathcal{A})$ converges to 1 for any $\delta > 0$. For the second condition $\mathcal{B}$, which holds w.p. $o(1)$ as $n \to \infty$, we must ensure that $E[f(N(n))|\mathcal{B})P(\mathcal{B})$ converges to zero. This trivially holds since $|f(N(n))| \leq E[N(n)]$ and Chernoff bounds produce an exponentially decaying tail for $P(\mathcal{B})$. Repeating the same reasoning with $f(N(n))^r$ for $r \geq 1$, we obtain convergence in r-th mean using Lemma 4. Applying this result to (152), we obtain $E[B_n] \to \nu t$ where the individual steps are shown earlier in (150).

Next, notice that B_n is a sum of dependent, but identically distributed, variables $\{-E[N] \log(1 - p_{ij})\}_{i,j}$. We next prove that $Var[-E[N] \log(1 - p_{ij})]$ decays to zero. First, notice that $X_n \to c < \infty$ in mean-square implies that $Var[X_n] \to 0$. Second, using the fact that $f(N(n)) \to 1$ in r-th mean, observe that $-E[N] \log(1 - p_{ij})$ converges to a constant in r-th mean for all $r \geq 1$, which gives us the desired result.

Now, for identically-distributed variables $\{X_i\}$, $Var[\sum_{i=1}^{n} X_i] \leq n^2 Var[X_1]$ and

therefore for any r.v. Y

$$Var\Big[\sum_{i=1}^{Y} X_i\Big] = E\Big[Var\Big[\sum_{i=1}^{Y} X_i|Y\Big]\Big] + Var\Big[E\Big[\sum_{i=1}^{Y} X_i|Y\Big]\Big]$$
$$\leq E[Y^2]Var[X_1] + Var[Y]E^2[X_1]. \tag{154}$$

Applying this result to (152) and noting that $\{W_i(t)\}_{i=1}^{n}$ are pairwise independent variables for $n \to \infty$, we get

$$Var[B_n] \leq \frac{n\Big(E[W_i(t)^2]\epsilon_n + Var[W_i(t)]\zeta_n\Big)}{E^2[N(n)]}, \tag{155}$$

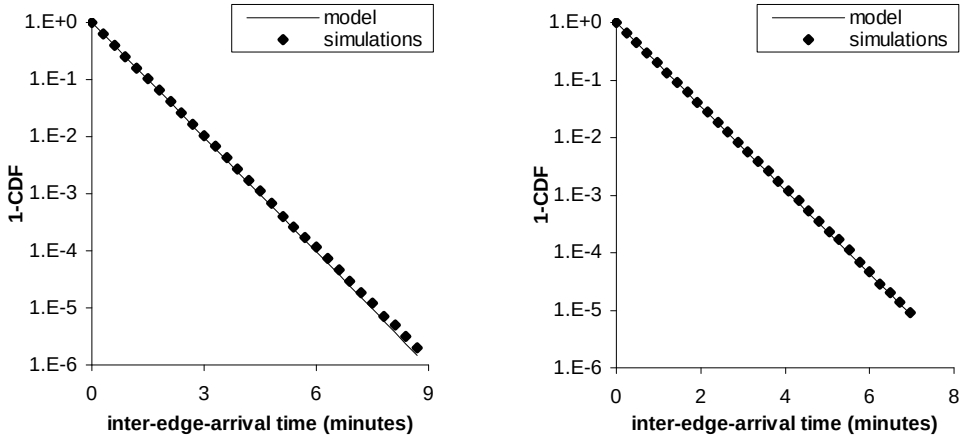
where $\epsilon_n = Var[-E[N]\log(1-p_{ij})]$ and $\zeta_n = E^2[-E[N]\log(1-p_{ij})]$. Observe that $n/E^2[N] \to 0$, $\epsilon_n \to 0$, and $\zeta_n \to 1$ as $n \to \infty$. Using similar arguments as in Lemma 13, it is easy to show that $E[W_i(t)^2]$ and $Var[W_i(t)]$ are both uniformly bounded in n. We then obtain that $Var[B_n] \to 0$ as $n \to \infty$.

Using Chebyshev's inequality, we get $B_n \to \nu t$ in probability. Finally, noticing that e^{-B_n} is uniformly integrable since it is always bounded in $[0, 1]$ as it represents the probability in (151), we obtain the desired convergence again following the reasoning in Lemma 4. □

5.2.4 Simulations

Fig. 17 shows the distribution of edge inter-arrival delays to a single node obtained in simulations with two types of systems. Notice in the sub-figures that the distribution of inter-arrival delay is nearly exponential with the rate given by (147). Additionally, Fig. 18 shows that the distribution of the number of edge arrivals to a node in an interval of size Δt approaches a Poisson distribution with the same rate ν in (147).

Finally, note that the Poisson result in Theorem 10 is not an *assumption* of the chapter as in prior work [39], [50], [61], but rather a *consequence* of the churn model



(a) system $\mathcal{H}$ with $n = 1000$ (b) system $\mathcal{VH}$ with $n = 5000$

Fig. 17. Distribution of edge inter-arrival delays approaches exponential with rate ν in (147) for $k = 10$ and $\theta = 10$ using 10^9 iterations.

introduced earlier.

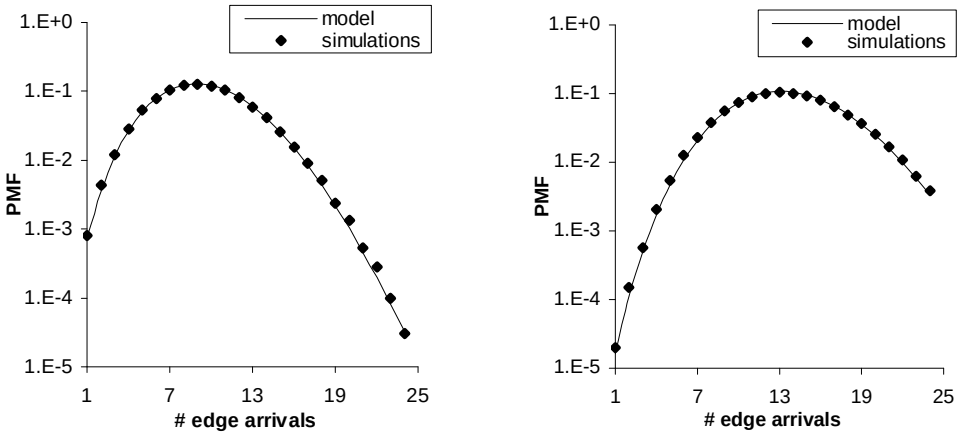
5.3. In-Degree

We now focus on understanding how the in-degree of each live user changes with time. For the rest of the chapter, we assume $n \to \infty$ and the edge arrival process to individual peers is Poisson with rate ν in (147).

5.3.1 Expected In-Degree

In a stationary system, define $X_n(t)$ to be the in-degree of a random online user at age $t \geq 0$. In this section, we focus on transient and limiting distributions of $X_n(t)$ under uniform selection of neighbors. We then have the following result.

Theorem 11. *Let $\{U(s)\}_{s\geq 0}$ be a pure renewal process with cycle length $R \sim H(x)$. Given that a user is alive in the system, its expected in-degree at fixed age $t \geq 0$*



(a) system $\mathcal{H}$ with $\Delta t = 6$ min (b) system $\mathcal{VH}$ with $\Delta t = 9$ min

Fig. 18. Distribution of the number of edge arrivals to a node in the interval $[t, t+\Delta t]$ in a system with $n = 1000$ users, $k = 10$, and $\theta = 10$. The lines show Poisson fits with ν in (147) at $t = 500$ hours and after 10^5 iterations.

converges as $n \to \infty$ *to a monotonically increasing function of age*

$$E[X_n(t)] \to k \int_0^\infty \left(E[U(x)] - E[U(x-t)]\right) H(dx), \tag{156}$$

where $E[U(x)] = \sum_{r=0}^{\infty} H^{*r}(x)$ *and* $E[U(x)] = 0$ *for* $x < 0$.

Proof. Fix $t > 0$, assume user 0 begins an ON period at time 0, and let i be any other alive user in its stationary state, which implies that its age at t follows $A_i(t) \sim H_i(x)$. Define $\tau = \max(t - A_i(t), 0)$ to be time from which both users are simultaneously present online and $I_i^c(t)$ to be an indicator variable of the event that i delivers an edge to user 0 in $[\tau, t]$ using its link c. Then, we are interested in computing

$$q_i := P(I_i^c(t) = 1 | Z_i(t) = 1, L_v > t, N). \tag{157}$$

Suppose $I_i^{cr}(t)$ is the indicator of user i hitting user 0 with an edge during its r-th attempt. Then, $q_i = \sum_{r=1}^{\infty} P(I_i^{cr}(t) = 1 | Z_i(t) = 1)$, where multiple edges from i

to user 0 occur w.p. 0 as $n \to \infty$. Observe that

$$P(I_i^{cr}(t) = 1 | Z_i(t) = 1) = \frac{P(U(A_i(t) - t) < r \leq U(A_i(t)))}{N(n)}$$

where $U(s)$ is the number of edges generated by i by time s along link c (i.e., number of renewals of a pure renewal process with cycle length R). Simplifying this expression:

$$q_i = \frac{1}{N(n)} E[U(A_i(t)) - U(A_i(t) - t) \mid Z_i = 1]. \tag{158}$$

We now arrive at the expected number of edges received by user 0 in $[0, t]$ from the entire system

$$\begin{aligned} E[X_n(t)] &= k \sum_{i=1}^{n} a_i E[q_i] \\ &= E\Big[\frac{k}{N(n)}\Big] \sum_{i=1}^{n} a_i \int_0^{\infty} E[U(x)] - E[U(x-t)] H_i(dx) \\ &\xrightarrow{a.s.} k \int_0^{\infty} \left(E[U(x)] - E[U(x-t)]\right) H(dx), \end{aligned} \tag{159}$$

which is the desired result. □

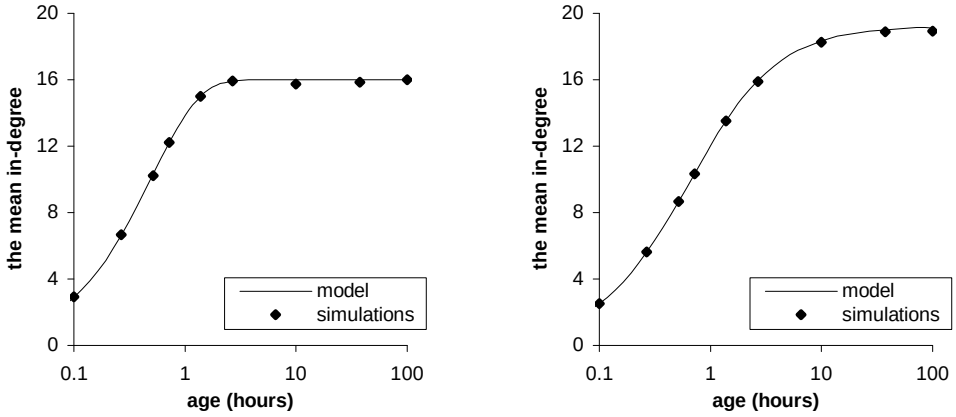
Model (156) can be written as

$$\lim_{n \to \infty} E[X_n(t)] = k \left(E[U(R)] - E[U(R-t)]\right), \tag{160}$$

where $U(R)$ is the number of renewals of process $\{U(s)\}_{s \geq 0}$ in a random interval $[0, R]$, where $R \sim H(x)$. Furthermore, as $t \to \infty$, (160) tends to $kE[U(R)]$, which provides a simple upper-bound at which the in-degree of each user saturates.

We next show that (160) can be expressed in simple closed-form for exponential lifetimes.

Theorem 12. *For exponential lifetimes L and $n \to \infty$, the mean in-degree at failure*



(a) exponential (161) (b) Pareto (156) with $\alpha = 3$

Fig. 19. Comparison of the model for $E[X(t)]$ to simulation results for $n = 2000$, $E[L] = 0.5$ hours, and $k = 8$ after 10^6 iterations.

$\theta = k$ *and*

$$E[X_n(t)] \to 2k(1 - e^{-t/E[L]}). \tag{161}$$

Proof. Since $F(x)$ is exponential, we have $H(x) = 1 - e^{x/E[L]}$. Then, the pure renewal process $\{U(s)\}$ with cycle length $R \sim H(x)$ is a Poisson process with a point at time 0. This leads to $E[U(x)] = 1 + x/E[L]$. Then, we have

$$\begin{aligned} E[X_n(t)] &\to k\Big(\int_0^t \big(1 + \frac{x}{E[L]}\big)\, H(dx) + \int_t^\infty \frac{x}{E[L]}\, H(dx)\Big) \\ &= k\Big(H(x) + \frac{E[\min(L,t)]}{E[L]}\Big) = 2k(1 - e^{-t/E[L]}). \end{aligned}$$

Finally, $\theta = \lim_{n\to\infty} \int_0^\infty E[X_n(t)]F(dt) = k$, which completes this proof. □

In (161), the mean in-degree of a node increases monotonically from $X_n(0) = 0$ when it arrives into the system to $E[X_n(\infty)] = 2k$ when its age tends to infinity. For the exponential case we directly use (161), while for the Pareto case we numerically solve (160). Simulation results in Fig. 19 demonstrate that the models are very accurate and indeed saturate at predicted values $2k$ and $kE[U(R)]$ as age $t \to \infty$.

Furthermore, if a node survives for more than 1 hour in the system, it develops an average of $12 - 15$ in-degree neighbors (depending on the distribution of L) and is unlikely to be isolated from the graph from that point on. It is also interesting to observe in the figure that the Pareto curve increases slower, but saturates at larger values, which suggests more resilience support for users with very large lifetimes. The saturation effect illustrated in Fig. 19 also shows that P2P implementations should cap user in-degree at values no smaller than the limit of (156) for $t \to \infty$. The corresponding upper bound in Gnutella (i.e., 30 in-degree neighbors) satisfies this condition for the two examples shown above.

5.4. Joint In/Out-Degree Model

Analytical results in the previous section show that the early stage in a node's life in the network is actually risky from the isolation point of view as it must rely solely on its out-degree neighbors. However, once a node survives this early stage, it increases its resilience to isolation through constantly arriving incoming edges. In this section, we combine the in-degree and out-degree models to derive the *joint* isolation probability of an arriving user. We drop subscript n and assume $n \to \infty$.

5.4.1 Preliminaries

Denote by $X^*(t)$ the out-degree of a node v at given age t and define it to be *isolated* when its in-degree and out-degree are simultaneously zero. Define *time to isolation* T to be the first-hitting time of both processes to state 0:

$$T = \inf\{t > 0 : X^*(t) = X(t) = 0 | X^*(0) = k, X(0) = 0\}. \tag{162}$$

Then the probability of node isolation is simply $\phi = P(T < L)$, where L is the

random lifetime of node v. Unlike in the out-degree process, a node does not replace its in-coming edges, which means that the in-degree and out-degree processes are independent of each other.

In the next subsections, we derive ϕ for systems with exponential user lifetimes and exponential search delays using two methods. The first approach provides an exact model using matrix algebra, while the second one shows an asymptotically accurate approximation that is available in simple closed-form.

5.4.2 Exponential Lifetimes (Exact Model)

Let pair $(X^*(t), X(t))$ be the joint process of out-degree and in-degree of a node at age t and (i, j) denote any admissible state of the joint process for $0 \leq i \leq k$ and $0 \leq j < n$. Recall that edge arrival at any node occurs according to a Poisson process with rate (147). Therefore, under uniform selection, incoming neighbors arrive to v at rate:

$$\nu = \frac{k + \theta}{E[L]} = \frac{2k}{E[L]} \tag{163}$$

since $\theta = k$ for exponential lifetimes. The current in-degree neighbors of v fail at rate $\mu = 1/E[L]$ due to the memoryless property of exponentials. This directly leads to the next result.

Lemma 14. *Given $L \sim \exp(\mu)$ and search times $S \sim \exp(\sigma)$ for finding replacement neighbors, the joint process $\{(X^*(t), X(t))\}$ is a homogeneous continuous-time Markov*

chain with a transition rate matrix $Q = (q_{uu'})$*:*

$$q_{uu'} = \begin{cases} i\mu & (i,j) \to (i-1,j) \\ (k-i)\sigma & (i,j) \to (i+1,j), \ \textit{for } i < k \\ j\mu & (i,j) \to (i,j-1) \\ 2k\mu & (i,j) \to (i,j+1) \\ -\Lambda_{ij} & (i,j) \to (i,j) \\ 0 & \text{otherwise} \end{cases}, \tag{164}$$

where u *and* u' *represent any suitable states of the joint chain satisfying transition requirements on the right side of (164) and* $\Lambda_{ij} = i\mu + (k-i)\sigma + j\mu + 2k\mu$*.*

Proof. Observe that given state $(X^*(t) = i, X(t) = j)$, there currently exist i out-going edges, $k-i$ searches in process, and j in-coming edges, and each is independent of one another. Since the in-coming edge arrival approaches a Poisson process at rate $2k\mu$ (see (163)), edges are $\exp(\mu)$ and search processes are $\exp(\sigma)$, the sojourn time in state (i,j) is thus exponential with rate:

$$\Lambda_{ij} = i\mu + (k-i)\sigma + j\mu + 2k\mu, \tag{165}$$

where the first two terms come from the out-degree process $W(t)$ and the last two from the in-degree process $X(t)$.

Denote by $p_{uu'}$ the probability that state u' is visited after some sojourn time in the current state u. Recall that when an out-going edge dies, a search starts immediately and its properties are independent of those of other search processes, edge lifetimes and the in-coming edge arrival process. This type of transition reduces $W(t)$ by 1 (and meanwhile increases the number of search processes by 1) in response to one failure of v's out-going edges, which is equivalent to the jump: $(i,j) \to (i-1,j)$

for $i > 0$. The corresponding probability that an out-going edge dies before any other event happens is $p_{(i,j)(i-1,j)} = i\mu/\Lambda_{ij}$.

Similarly, the second type of transition as a result of finding a replacement neighbor is written as $(i,j) \to (i+1,j)$ for $i < k$. Its related probability is $p_{(i,j)(i+1,j)} = (k-i)\sigma/\Lambda_{ij}$. The third type of transition responding to one failure of existing in-incoming edges is denoted by $(i,j) \to (i,j-1)$ for $j > 0$, and the transition probability is $p_{(i,j)(i,j-1)} = j\mu/\Lambda_{ij}$. Finally, the last type of transition caused by the arrival of a new in-coming edge is a jump: $(i,j) \to (i,j+1)$ for $j < n-1$ with probability $p_{(i,j)(i,j+1)} = 2k\mu/\Lambda_{ij}$.

By recognizing that the jumps behave like a discrete-time Markov chain and the sojourn times in each state are independent exponential random variables, we immediately conclude that the joint chain $\{(X^*(t), X(t))\}$ is a homogeneous continuous-time Markov chain with a transition rate matrix $Q = (q_{uu'})$ shown in (164) . □

It is convenient to treat $\{(X^*(t), X(t))\}$ as an absorbing Markov chain in order to derive the PDF of the first-hitting time T on state $(0,0)$. Assuming $(0,0)$ is an absorbing state, we can write Q in canonical form as:

$$Q = \begin{pmatrix} 0 & 0 \\ \mathbf{r} & Q_0 \end{pmatrix}, \tag{166}$$

where Q_0 is the rate matrix obtained by removing the rows and columns corresponding to state $(0,0)$ from Q and $\mathbf{r}$ is a column vector of transition rates into state $(0,0)$.

Applying Theorem 4 in Chapter IV, we obtain the next result.

Corollary 4. *For exponential lifetimes $L \sim \exp(\mu)$ and exponential search delays $S \sim \exp(\sigma)$, the probability of node isolation is given by:*

$$\phi = \pi(0) V B V^{-1} \mathbf{r}, \tag{167}$$

Table II. Exact model (167) and simulations ($n = 2000$, $E[L] = 0.5$ hours)

$E[S]$	$k = 6$		$k = 8$	
min	Simulations	Model (167)	Simulations	Model (167)
6	3.63×10^{-6}	3.61×10^{-6}	2.80×10^{-8}	2.87×10^{-8}
18	3.15×10^{-5}	3.17×10^{-5}	5.91×10^{-7}	5.98×10^{-7}
30	6.04×10^{-5}	6.08×10^{-5}	1.48×10^{-6}	1.46×10^{-6}
42	8.38×10^{-5}	8.37×10^{-5}	2.30×10^{-6}	2.27×10^{-6}
60	1.06×10^{-4}	1.09×10^{-4}	3.27×10^{-6}	3.28×10^{-6}

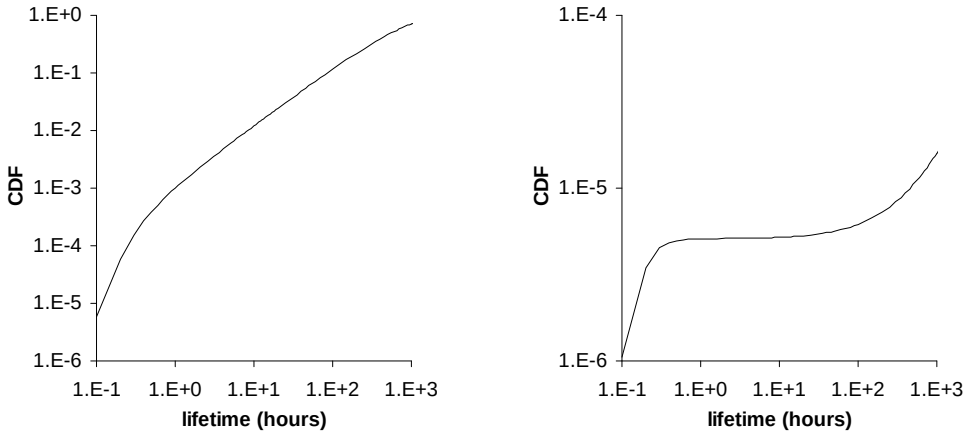
where V is a matrix of eigenvectors of Q_0, $B = diag(b_j)$ is a diagonal matrix with $b_j = 1/(\mu - \xi_j)$, ξ_j is the j-th eigenvalue of Q_0, and $\pi(0) = (\pi_{(i,j)}(0))$ is the initial state distribution with $\pi_{(k,0)}(0) = 1$ and $\pi_{(i,j)}(0) = 0$ in all other pairs.

We verify (167) in simulations shown in Table II, which shows that our results are indeed very accurate. While (167) provides values ϕ that are smaller than isolation probability ϕ_{out} of the out-degree model [43] by several orders of magnitude, it is still unclear what impact in-degree has on the probability that a user gets isolated as its age increases and how large the improvement ratio ϕ_{out}/ϕ is. We study these issues below.

5.4.3 Isolation with Increased Age

To better understand the impact of in-degree on ϕ, let us define the first hitting time T_{out} on state 0 of the out-degree process $\{X^*(t)\}$, i.e., $T_{out} = \inf\{t > 0 : X^*(t) = 0 | X^*(0) = k\}$. Analysis in [43] shows that $\{X^*(t)\}$ is a birth-death Markov chain and derives its CDF function $P(T_{out} < t)$ in matrix form. We use this result and the CDF of T derived in the proof of Theorem 4 to compare the distribution of isolation times in the joint in/out degree model with that studied in [43]. We plot the exact

distributions of both T_{out} and T as functions of user age in Fig. 20. Notice in the figure that $P(T_{out} < t)$ increases almost linearly in time t indicating that users with large lifetimes have proportionally higher probabilities of isolation. In contrast, the curve of $P(T < t)$ becomes almost flat as time t increases beyond 0.5 hours showing that users with lifetimes in the range $[0.5, 200]$ hours exhibit almost the same isolation probabilities. In fact, once the initial 1/2-hour period is over, isolation probability is orders of magnitude smaller than in the initial phase. As user age increases above 200 hours, the CDF of T slowly increases in time since $X(t)$ becomes saturated and can no longer keep up with the increased possibility of neighbor failure.



(a) T_{out} (b) T

Fig. 20. The CDF of T_{out} and T for exponential lifetimes with $E[L] = 0.5$ hours, exponential search delays with $E[S] = 0.1$ hours, and $k = 6$.

5.4.4 Exponential Lifetimes (Asymptotic Model)

Although (167) provides exact results for ϕ, it relies on numerical matrix algebra. Our next task is to obtain a simple closed-form solution to ϕ when the mean search delay $E[S] \to 0$.

We begin with obtaining the asymptotic distribution of the first-hitting time

T_{out} onto state 0 of the out-degree process $\{X^*(t)\}$ and then obtain node isolation probability ϕ_{out} which only considers out-degree.

Lemma 15. *For $L \sim \exp(\mu)$ and $S \sim \exp(\sigma)$, ϕ_{out} converges to the following as $E[S] \to 0$:*

$$\phi_{out} = \rho k/(1+\rho)^k, \tag{168}$$

where $\sigma/\mu = E[L]/E[S]$.

Proof. Using previous results in [4], we know that for Markov chain $\{X^*(t)\}$, the first hitting time T_{out} of a rare event (e.g., state 0 of $\{X^*(t)\}$) behaves as an exponential random variable with rate $1/E[T_{out}]$:

$$P(T_{out} < t) = 1 - e^{-t/E[T_{out}]}, \text{ as } E[S] \to 0, \tag{169}$$

where $E[T_{out}]$ is available in closed form [43]:

$$E[T_{out}] = \frac{E[S]}{k}(1+\rho)^k, \tag{170}$$

where S denotes the search delay and $\rho = E[L]/E[S]$. Observe that $E[T_{out}] \to \infty$ as $E[S] \to 0$. Thus by Taylor expansion, (169) reduces to:

$$P(T_{out} < t) = t/E[T_{out}], \text{ as } E[S] \to 0, \tag{171}$$

showing that asymptotically T_{out} behaves like a uniform random variable. Taking the derivative of (171), we obtain the asymptotic result on the PDF of T_{out}:

$$f_{T_{out}}(t) = 1/E[T_{out}], \text{ as } E[S] \to 0. \tag{172}$$

It is then straightforward to obtain:

$$\phi_{out} = P(T_{out} < L) = \int_0^\infty P(L > t) f_{T_{out}}(t) dt = \frac{E[L]}{E[T_{out}]}, \tag{173}$$

as $E[S] \to 0$, which is the desired result. $\square$

We then derive the asymptotic CDF of T of the joint chain $\{X^*(t), X(t)\}$ in the following.

Lemma 16. *Given $L \sim \exp(\mu)$ and $S \sim \exp(\sigma)$, the CDF of T onto state $(0,0)$ of the joint in/out-degree process approaches the following as $E[S] \to 0$:*

$$P(T < x) = e^{-2k}(Ei(2k) - Ei(2ke^{-\mu x}))\phi_{out}, \tag{174}$$

where ϕ_{out} is given by (168) and $Ei(x) = -\int_{-x}^{\infty} e^{-z}/z dz$ is the exponential integral.

Proof. Observe that user lifetime L is small compared to the value of the first hitting time T on state $(0,0)$. Therefore, $P(T < L)$ is mainly affected by the CDF $P(T < x)$ only for small x. Next, note that the probability that out-degree process $\{X^*(t)\}$ hits more than once on state 0 within interval $[0, x]$ for small x is negligible when $E[S] \to 0$ (i.e., $E[T_{out}] \to \infty$). Thus, based on the property of the first hitting time T_{out} and the probability that the in-degree is zero at epoch T_{out}, we obtain a simple formula for the asymptotic CDF of T:

$$P(T < x) = \int_0^x P(X(t) = 0) f_{T_{out}}(t) dt, \tag{175}$$

as $E[S] \to 0$, where $P(X(t) = 0)$ is given in (161) for exponential lifetimes. Substituting (161) and (172) into (175) leads to the following as $E[S] \to 0$:

$$\begin{aligned} P(T < x) &= \frac{1}{E[T_{out}]} \int_0^x e^{-2k(1-e^{-\mu t})} dt \\ &= \frac{e^{-2k}}{\mu E[T_{out}]} \int_{-2ke^{-\mu x}}^{-2k} \frac{e^{-z}}{z} dz. \end{aligned} \tag{176}$$

Notice that:

$$\int_{-2ke^{-\mu x}}^{-2k} \frac{e^{-z}}{z} dz = \int_{-2ke^{-\mu x}}^{\infty} \frac{e^{-z}}{z} dz - \int_{-2k}^{\infty} \frac{e^{-z}}{z} dz. \tag{177}$$

Substituting (177) into (176) and using $\mu = 1/E[L]$ and (168), we easily establish (174). □

The asymptotic result on the CDF of T for $E[S] \to 0$ immediately leads to finding isolation probability ϕ, as shown next.

Theorem 13. *For $L \sim \exp(\mu)$ and $S \sim \exp(\sigma)$, isolation probability converges to the following as $E[S] \to 0$:*

$$\phi = \frac{1 - e^{-2k}}{2k} \phi_{out}, \tag{178}$$

where $\phi_{out} = \rho k/(1+\rho)^k$ and $\rho = \sigma/\mu = E[L]/E[S]$.

Proof. Integrating (174) using the PDF $f(x) = \mu e^{-\mu x}$ of user lifetimes, we have:

$$\begin{aligned} \phi &= \int_0^\infty P(T < x) f(x) dx \\ &= e^{-2k} \Big(\mathrm{Ei}(2k) - \int_0^\infty \mathrm{Ei}(2ke^{-\mu x})) f(x) dx \Big) \phi_{out} \\ &= e^{-2k} \Big(\mathrm{Ei}(2k) - \frac{1}{2k} \int_0^{2k} \mathrm{Ei}(x) dx \Big) \phi_{out}. \end{aligned} \tag{179}$$

Observe that:

$$\int_0^{2k} \mathrm{Ei}(x) dx = 1 - e^{2k} + 2k\mathrm{Ei}(2k). \tag{180}$$

Substituting (180) into (179), we easily obtain (178). □

It can be seen from (178) that by considering both in-degree and out-degree, the probability of node isolation is reduced by a factor of approximately $2k$ for non-trivial k. The reason for this relatively small improvement is that only a handful of users benefit from the in-degree in their isolation resilience since the majority of users depart very quickly and are unable to accumulate any in-degree neighbors. Nevertheless, analysis of this section has important consequences as it shows that *the most reliable users of the system (i.e., those with large lifetimes) extract huge*

Table III. Convergence of (178) to (167) for $E[L] = 0.5$ Hours and $k = 6$

$E[S]$	Exact model (167)	Approx. model (178)	Relative error
36 sec	8.721×10^{-10}	1.421×10^{-9}	62.91%
3.6 sec	1.498×10^{-14}	1.581×10^{-14}	5.57%
360 ms	1.589×10^{-19}	1.598×10^{-19}	0.55%
36 ms	1.600×10^{-24}	1.600×10^{-24}	0

benefits from the in-degree process and are thus allowed to continue providing services to others with much higher probability than possible with just the out-degree.

To complete this section, Table III shows the relative approximation error of (178) and confirms its asymptotic accuracy. For large S, our numerical results from the exact model suggest that (178) provides an upper bound on the isolation probability, where ϕ_{out}/ϕ is 3-10 times larger than the $2k$ suggested by (178). For instance, for fixed $E[L] = 0.5$ hours and $k = 6$, ratio ϕ_{out}/ϕ is 39 when $E[S] = 2$ minutes and 120 when $E[S] = 6$ minutes.

5.5. Summary

This chapter formally proved that the edge-arrival process to each user under uniform selection approached Poisson as system size became sufficiently large. We then developed numerous closed-form results describing transient in-degree distribution and isolation probability under the joint in/out degree model.

CHAPTER VI

LINK LIFETIMES IN DHTS

6.1. Introduction

Traditional metrics in analysis of resilience of P2P systems have been the ability of the graph to stay connected during user departure [45], [50], [61], behavior of immediate neighbors during churn [39], data delivery ratio [84], evolution of out-degree [42] and in-degree [93], and churn rate in the set of participating nodes [26]. All metrics above depend on one fundamental parameter of churn – *link lifetime*, which is defined as the delay between formation of a link and its disconnection due to a sudden departure of the adjacent neighbor.

Recall that in many P2P networks, each joining user v creates and monitors k links to other peers. Link behavior is often modeled as an ON/OFF process [43] in which each link is either ON at time t, which means that the corresponding user is currently alive, or OFF, which means that the user adjacent to the link has failed and its failure is in the process of being detected and repaired. ON durations of links are *link lifetimes* and their OFF durations are *repair delays*.

If links do not switch to other users during each ON duration (i.e., keep connecting to the same neighbors until they fail), then link durations are simply *residual lifetimes* of original neighbors. We call this model *non-switching* and note that it applies to certain unstructured P2P networks [25] and some DHTs [58]. Link lifetimes for non-switching systems have been studied in fair detail under both age-independent [42] and age-biased [84] selection. However, many DHTs actively switch links to new neighbors before the current neighbor dies in order to balance the load and ensure

DHT consistency. We call such systems *switching* and note that their link lifetimes require entirely different modeling techniques, which we present below (in the notation of [26], switching/non-switching are agnostic neighbor replacement strategies, where the former is called Active Preference List (APL) and the latter encompasses both Passive Preference List (PPL) and Random Replacement (RR)).

6.1.1 Analysis of Existing DHTs

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

We next obtain the Laplace transform of the distribution of R and derive its expected value $E[R]$ for general user lifetimes L, including heavy-tailed cases. We then use this result to show that in systems with exponential peer lifetimes, link lifetime R follows the same exponential distribution, which indicates that for such cases link lifetimes are very similar to those in networks without switching [42]. However, for heavy-tailed peer lifetimes (e.g., Pareto) observed in many real P2P networks [12], [74], [89], our model shows that R is stochastically *smaller* than the residual lifetime Z of the initial neighbor holding the link and, as first observed in [27], the mean link lifetime $E[R]$ is very close to $E[L]$. This is in stark contrast to the results of [42] where $E[R]$ is several times larger than $E[L]$ depending on Pareto shape α of the lifetime distribution (e.g., $E[R] \approx 11.1E[L]$ for $\alpha = 1.09$ observed in [89] and $E[R] \approx 16.6E[L]$ for $\alpha = 1.06$ observed in [12]). This phenomenon occurs because older (i.e., more

reliable) neighbors in DHTs are replaced with new arrivals that exhibit much shorter remaining lifetimes. As a result, classical DHTs unfortunately do not extract any benefits from heavy-tailed user lifetimes and suffer much higher link churn rates than the corresponding unstructured systems [42]. A similar conclusion was obtained in [26] for query failure rates in Chord.

6.1.2 Improvements

One method of overcoming the problem identified above is to utilize randomized DHTs (e.g., randomized Chord [29], randomized hypercube [56], and Symphony [55]) in which the i-th finger pointer of a given user v is randomly selected from some set S_i of possible locations in the DHT space. By trying multiple options in S_i and linking to the user with the best characteristics (which we determine below), the hope is to improve link lifetime and reduce the impact of churn on system performance. Note that this method only works when set S_i is sufficiently large. We assume that each node has at least one link that satisfies this condition. The first randomized technique, which we call *max-age*, selects m points in S_i uniformly randomly and connects v to the user with the largest age (this method was suggested in [84] for DHTs and [96] for unstructured P2P systems). While quite effective in non-switching scenarios, this strategy has minimal impact in DHTs since link lifetime is determined by the remaining session length of not the *first*, but the *last* neighbor holding the link.

To overcome this limitation, we propose a novel randomized strategy that stems from our model of link lifetime R. Our theoretical results show that neighbors with larger zones (e.g., in Chord [79], with larger distance to the predecessor) are less reliable as they are more likely to be hit by a new arrival whose remaining lifetime will be small. To extract benefits from randomized selection, we show that users must

prefer neighbors in S_i with the *smallest zone size* rather than maximum age or any other characteristic. We call this strategy *min-zone* and show that it is vastly more effective than max-age selection given lifetime distributions observed in real systems [12], [89]. In addition to reduced link churn, min-zone selection benefits DHTs by balancing the load such that users with smaller zone sizes are responsible for fewer keys while forwarding more queries.

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

We finish the chapter by showing that under min-zone selection and shape parameter $1 < \alpha \leq 2$, the mean link lifetime $E[R]$ tends to infinity as the number of samples m becomes large. We also suggest simple formulas for $E[R]$ using examples of Pareto shape α obtained from recent measurements [12], [89] and show simple results demonstrating the growth rate of $E[R]$ as a function of m.

6.2. General DHT Model

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

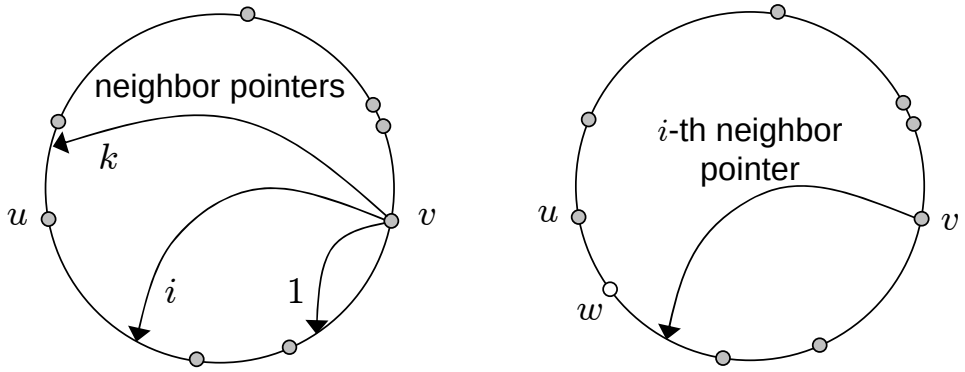
6.2.1 Assumptions

Without loss of generality, we assume that the network maps keys and users into the same identifier (ID) space, which is a continuous ring in the interval $[0, 1)$ [60]. Each user is responsible for a fraction of the DHT space from its predecessor to itself, which we call the user's zone. To facilitate routing, each joining peer v selects and then monitors using some stabilization technique k links in the DHT space as shown in Fig. 21(a).

For the churn model, we adopt the framework of n alternating renewal processes representing periodic online/offline behavior of users (see Chapter III) observed in real P2P systems [26], [89]. While the total number of users n is fixed, the number of *currently alive* peers N_t at time t is a random process that fluctuates over time. Once stationarity is reached, we usually replace N_t with its limiting version $N = \lim_{t\to\infty} N_t$. We finally assume that when a particular user rejoins the system, it generates a new random ID (e.g., based on its IP-port pair) instead of using the same fixed hash. Note that the use of new IDs helps balance the load in the DHT [79], [90]. As a consequence of this churn model [93, Theorem 5], user arrivals into the system follow a Poisson process with a constant rate $\lambda = E[N]/E[L]$, where $E[N]$ is the average number of users in the steady state and $E[L]$ is the mean user lifetime.

6.2.2 Neighbor Dynamics

Note that the main focus of the chapter is on the behavior of one particular link i in Fig. 21(a) (other links are similar) and the lifetimes of neighbors adjacent to it during v's online session. As user v continues to stay in the system, the identity of its neighbors (i.e., successors of its neighbor pointers) may change over time as users join and leave the system. There are two types of changes in neighbor tables – graceful

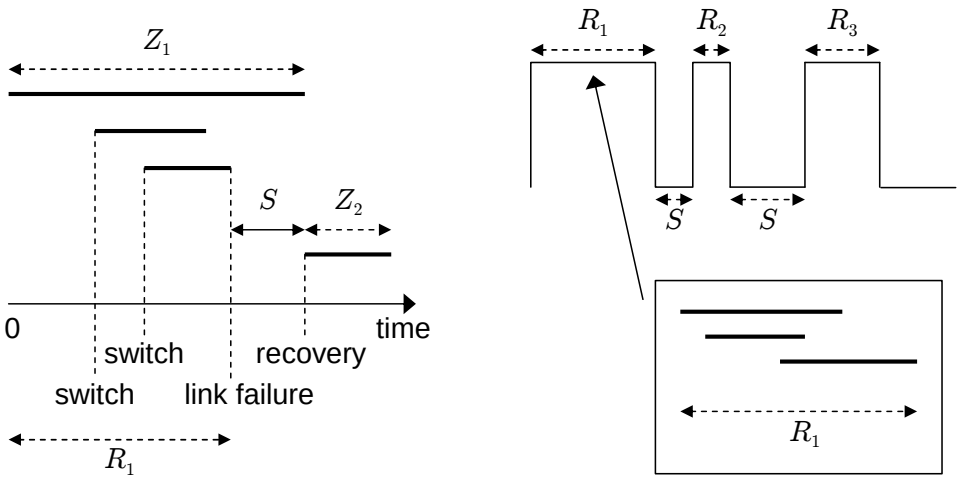


(a) k links of user v (b) arriving node w

Fig. 21. User v's neighbors in the DHT.

handoffs of an existing zone to another user and node departures without explicit notification of v [79]. The former type, which we call a *switch*, occurs when a new arrival takes ownership of a link by becoming the new successor of the corresponding neighbor pointer. This is shown in Fig. 21(b) where a new arrival w splits the zone of an existing neighbor u and becomes the new neighbor of v. The latter type of neighbor change, which we call a *recovery*, happens when an existing neighbor dies and the successor of the failed neighbor takes over that zone to become the new neighbor of v.

We next define several additional metrics to facilitate explanation in later parts of the chapter. Notice that one cycle in the life of a particular neighbor pointer is composed of several switches and one recovery as shown in Fig. 22(a). In the figure, thick horizonal lines represent online presence of peers that own v's neighbor pointer in the DHT space. The topmost line is the original neighbor with residual lifetime Z_1 acquired by v during join. As peers split the zone of the current neighbor, the link switches to two additional users. Switch is complete after a new user performs all join tasks [79]. Once the last user dies at time R_1, the link is considered dead and a replacement process is initiated. Specifics of detecting failure are not essential to



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

Fig. 22. The i-th link failure and replacement of user v who joins at time 0 in a DHT, $1 \leq i \leq k$.

our results as repair delay is not studied in this chapter. Recovery is finished after S time units when another node takes over the zone of the dead peer and is selected as v's new neighbor.

In all other aspects, the second recovery cycle behaves identical to the first one and leads to link failure after R_2 time units. This ON/OFF nature of the link process is shown in Fig. 22(b) where we assume that all repair delays S are *i.i.d.* random variables, but the distribution of link lifetimes $R_1, R_2, \ldots$ may depend on the cycle number (in fact they do in certain cases studied below).

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

butions.

6.3. Link Lifetime Model

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

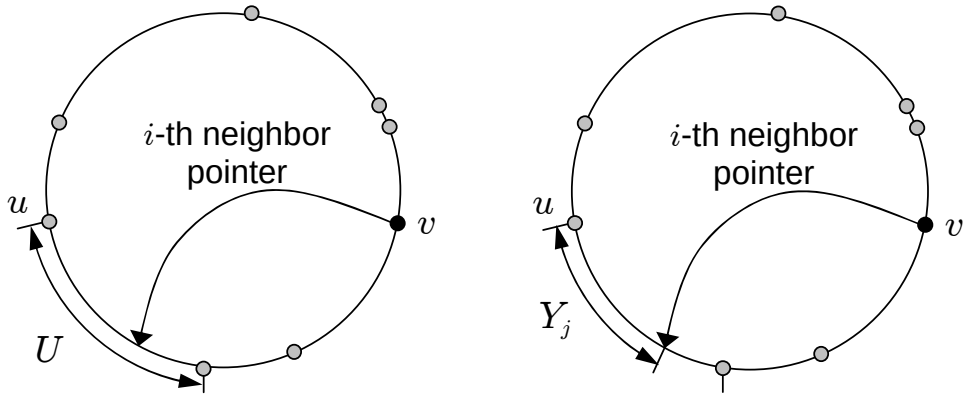
6.3.1 Preliminaries

Recall that arriving users split zones of existing nodes based on a uniformly random hashing function. Denote by U the random zone size of existing users in a stationary system as shown in Fig. 23(a). Further assume that during join or the current recovery step that starts cycle j, successor u takes over pointer i as shown in Fig. 23(b). Then, define Y_j to be the *remaining zone size* between this pointer and the index of u. Intuitively, if the remaining zone Y_j is large, then it is likely that a new arrival will soon split the zone and the ownership of the link will be transferred to another peer. Therefore, link lifetimes are determined not by the distribution of U, but rather by that of Y_j. We derive both metrics later in the chapter and next show how they can be used to obtain $R_1, R_2, \ldots$

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

$$P(R_j < x) = \int_0^\infty P(R(y) < x) f_{Y_j}(y) dy, \tag{181}$$

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



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

Fig. 23. Zone size U and remaining zone size Y_j of user u.

expectation of R_j as:

$$E[R_j] = \int_0^\infty E[R(y)] f_{Y_j}(y) dy. \tag{182}$$

Thus, the task of deriving link lifetime R_j is reduced to analyzing the properties of conditional link lifetime $R(y)$ and the distribution of remaining zone size Y_j. In the rest of this section, we construct a semi-Markov process for each $R(y)$ and leave the derivation of the distribution of Y_j for deterministic DHTs to Section 6.4. and that for randomized DHTs to Section 6.5.

6.3.2 Neighbor Dynamics

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

Then, it is easy to see that $\{A_\delta^y; \delta \geq 0\}$ is a continuous-time stochastic process

with state space $\{F, 0, 1, 2, \ldots\}$ whose state transitions are shown in Fig. 24. As depicted in this figure, for each state $i \geq 0$, the process can jump into either state $i+1$, which means that a given zone is further split by a new arrival (i.e., the number of switches increases by 1), or state F, which represents link failure. The initial state of the process at time 0 is always 0.

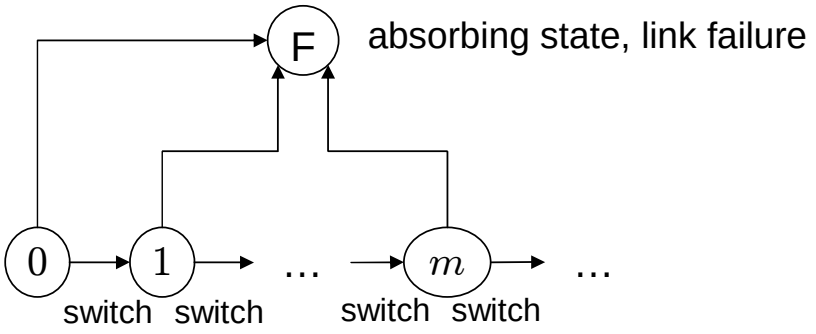


Fig. 24. State diagram for the process $\{A_\delta^j, \delta \geq 0\}$ of neighbor changes.

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

$$R(y) = \inf\{\delta > 0 : A_\delta^y = F | A_0^y = 0, Y_j = y\}. \tag{183}$$

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

Theorem 14. *Process $\{A_\delta^y, \delta \geq 0\}$ for a given remaining zone size $Y_j = y$ is a regular semi-Markov chain. The sojourn time τ_i in state i follows the following general distribution:*

$$P(\tau_i > x) = \begin{cases} P(W_0 > x)P(Z_j > x) & i = 0 \\ P(W_i > x)P(L > x) & i \geq 1 \end{cases}, \tag{184}$$

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

$$\lambda_i = \frac{E[N]y}{E[L]2^i}, \quad i \geq 0, \tag{185}$$

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

$$p_{i,i+1} = \begin{cases} P(W_0 < Z_j) & i = 0 \\ P(W_i < L) & i \geq 1 \end{cases}, \tag{186}$$

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

Proof. For the heterogeneous churn model of [93] used in this work, new user arrivals into the DHT space approach a Poisson process with constant rate [93, Theorem 5]:

$$\lambda = \frac{E[N]}{E[L]}, \tag{187}$$

where $E[N]$ is the mean number of users in an equilibrium system and $E[L]$ is the mean user lifetime. Then from the Marked Poisson theorem [70], the arrival process into any fixed zone with size y is Poisson with average rate:

$$\lambda_0 = \lambda p_y, \tag{188}$$

where $p_y = y$ is the probability that a given zone of length y is selected from the DHT space $[0, 1)$.

Next, observe that the wait time W_0 to transition from state 0 to state 1 (i.e., the delay before the next arrival into the remaining zone of size y between the neighbor pointer and the current neighbor) is exponentially distributed as $\exp(\lambda_0)$. As the given zone is successively divided by new arrivals over time, its length is reduced over

time, which in turn reduces the user arrival rate into the zone. Since a given zone of length y is uniformly divided under random split by a new arrival, the expected length of the new zone is simply $y/2$. This implies that the wait time W_i to transition from state i to state $i+1$ is exponential with rate:

$$\lambda_i = \frac{\lambda_0}{2^i} = \frac{E[N]y}{E[L]2^i}, \quad i \geq 0, \tag{189}$$

which depends not only on state i, but also the initial zone size y.

We now consider transitions to state F. Given $A_\delta = i$, $i \geq 1$, a jump to state F is triggered by the departure of the current user, which happens L time units after the chain arrives to state i, where L is the random user lifetime. For state $i = 0$, the delay before the jump to state F is slightly different and equals the original user's remaining lifetime Z_j where j is the cycle number of R_j. It then follows that due to the independence among user departures and arrivals in a sufficiently large system, the sojourn time τ_i in state i is simply:

$$\tau_i = \begin{cases} \min(W_0, Z_j) & i = 0 \\ \min(W_i, L) & i \geq 1 \end{cases}, \tag{190}$$

where $W_i \sim \exp(\lambda_i)$ and is independent of Z_j and L. Since Z_j and L may follow general distributions, respectively, sojourn time τ_i may have a non-exponential distribution.

Finally, transition probability $p_{i,i+1}$ from state i to $i+1$ is given by:

$$p_{i,i+1} = \begin{cases} P(W_0 < Z_j) & i = 0 \\ P(W_i < L) & i \geq 1 \end{cases}, \tag{191}$$

and the probability $p_{i,F}$ to absorb from state i is equal to $1 - p_{i,i+1}$. Note that due to $W_i \to \infty$ for $i \to \infty$, it is clear that $p_{i,i+1} \to 0$ as $i \to \infty$ and the decay rate is

exponentially fast. Thus, $\{A_\delta^y\}$ is regular.

Recognizing that these transitions behave like a discrete-time Markov chain and sojourn times in states depend only on their current states and follow general distributions, we immediately conclude that $\{A_\delta^i\}$ is a regular semi-Markov chain (SMC). □

This theorem shows in (185) that as the number of switches within a zone (i.e., variable i) increases, arrival rate λ_i of news users into the zone decreases exponentially fast (or alternatively, the mean waiting time $E[W_i]$ until the next arrival increases at the same rate). As $i \to \infty$, the likelihood of a new arrival into the zone diminishes and the delay in state i becomes simply the lifetime of the last user holding the edge. For small i, however, analysis is much more complex as shown in the next subsection.

6.3.3 Conditional Link Lifetimes

Next, we study the distribution and expectation of conditional link lifetime $R(y)$. To understand our next theorem, several definitions are necessary. First, denote the CDF of sojourn time τ_i in state i by:

$$G_i(t) = P(\tau_i < t). \tag{192}$$

Second, observing from (184) that τ_i of chain $\{A_\delta^y\}$ is independent of the next state, define a semi-Markov kernel matrix $Q(t) = [q_{ik}(t)]$ using [14]:

$$q_{ik}(t) = p_{ik} G_i(t), \quad i, k \in \{F, 0, 1, \ldots\}, \tag{193}$$

where p_{ik} is the transition probability from state i to state k given in (186). The Laplace (Stieltjes) transform of $q_{ik}(t)$ is then simply:

$$\hat{q}_{ik}(s) = \int_0^\infty e^{-st} dq_{ik}(t) = p_{ik} \int_0^\infty e^{-st} dG_i(t). \tag{194}$$

Finally, define the Laplace transform of the first hitting time $R(y)$ from state 0 to F as:

$$\hat{R}(s,y) = E[e^{-sR(y)}]. \tag{195}$$

Although it is known that the Laplace transform of the first-hitting time of a semi-Markov chain can be computed using spectral properties of kernel $Q(t)$ [11], this approach hides the effect of system parameters on the resulting distribution. Due to the simplicity of state transitions of chain $\{A_\delta^y\}$, we next derive $\hat{R}(s,y)$ without involving matrix operations on $Q(t)$.

Theorem 15. *The Laplace transform $\hat{R}(s,y)$ of conditional link lifetime $R(y)$ is given by:*

$$\hat{R}(s,y) = \hat{q}_{0F}(s) + \sum_{k=1}^{\infty} \left(\prod_{i=0}^{k-1} \hat{q}_{i,i+1}(s) \right) \hat{q}_{kF}(s), \tag{196}$$

where $\hat{q}_{ik}(s)$ are shown in (194)*.*

Proof. Generalize the first hitting time from any starting state $i \geq 0$ to state F as:

$$T_{iF} = \inf\{\delta > 0 : A_\delta^y = F | A_0^y = i, Y_j = y\} \tag{197}$$

and define the following Laplace transform for T_{iF}:

$$\hat{T}_{iF}(s) = E[e^{-sT_{iF}}] = \int_0^\infty e^{-st} dF_{T_{iF}}(t), \tag{198}$$

where $F_{T_{iF}}(t)$ is the CDF of T_{iF}. Then, from first-step analysis, (198) can be transformed into:

$$E[e^{-sT_{iF}}] = p_{iF} E[e^{-s\tau_i}] + p_{i,i+1} E[e^{-s(\tau_i + T_{i+1,F})}], \tag{199}$$

where p_{ik} is the transition probability from state i to k shown in (186). Noting that τ_i

is independent of $T_{i+1,F}$ and conditioning on the current state being i, (199) reduces to:

$$\begin{aligned} E[e^{-sT_{iF}}] &= p_{iF}E[e^{-s\tau_i}] + p_{i,i+1}E[e^{-s\tau_i}]E[e^{-sT_{i+1,F}}] \\ &= \hat{q}_{iF}(s) + \hat{q}_{i,i+1}(s)E[e^{-sT_{i+1,F}}], \end{aligned} \tag{200}$$

where $\hat{q}_{i,k}(s)$ is defined in (194). Using the above recurrent functions and observing that $\hat{q}_{i,i+1}(s) \to 0$ for $i \to \infty$ (due to transition probability $p_{i,i+1} \to 0$ in this case), we readily obtain:

$$E[e^{-sT_{0F}}] = \hat{q}_{0F}(s) + \sum_{k=1}^{\infty} \left(\prod_{i=0}^{k-1} \hat{q}_{i,i+1}(s) \right) \hat{q}_{kF}(s), \tag{201}$$

which establishes (196) upon recalling that $R(y)$ is defined as T_{0F}. □

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

Theorem 16. *The expected conditional link lifetime is:*

$$E[R(y)] = E[\tau_0] + \sum_{k=1}^{\infty} \left(\prod_{i=0}^{k-1} p_{i,i+1} \right) E[\tau_k], \tag{202}$$

where $E[\tau_k]$ is the expected sojourn time in state k shown in (184) *and $p_{i,i+1}$ are state transition probabilities in* (186).

Proof. Given that the chain currently is in state $i \geq 0$, it can jump either to state F or $i + 1$. Then by conditioning on the first jump, it is not hard to see that:

$$E[T_{iF}] = E[\tau_i] + p_{i,i+1}E[T_{i+1,F}], \tag{203}$$

where T_{iF} is defined in (197). Using the above recurrence functions, we easily obtain:

$$\begin{aligned} E[R(y)] &= E[T_{0F}] = E[\tau_0] + p_{01}E[T_{1F}] \\ &= E[\tau_0] + p_{01}\left(E[\tau_1] + p_{12}E[T_{2F}]\right) \\ &= E[\tau_0] + \sum_{k=1}^{\infty}\left(\prod_{i=0}^{k-1} p_{i,i+1}\right)E[\tau_k], \end{aligned} \tag{204}$$

where the last step is obtained by induction and recalling that $p_{i,i+1} \to 0$ for $i \to \infty$. □

Theorems 14–16 demonstrate that variable $R(y)$ is fully determined by user lifetimes L and residual neighbor lifetimes Z_j. Our remaining steps are to analyze the properties of Z_j and derive the distribution of remaining zone sizes Y_j for both deterministic and randomized DHTs.

6.4. Deterministic DHTs

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

6.4.1 Residual Lifetimes of Neighbors

Using the user churn model summarized in Section 6.2.1, it has been shown in Theorem 2 that the distribution of neighbor residual lifetime under uniform selection converges to the following equilibrium CDF as system age $t \to \infty$:

$$H(x) = \frac{1}{E[L]}\int_0^x (1 - F(u))du, \tag{205}$$

where $F(x)$ is the user lifetime distribution. Since recovery in our DHT model is not biased with respect to user age, (205) is also the CDF of residual lifetime for users that are found during recovery, which we formally state in the next lemma.

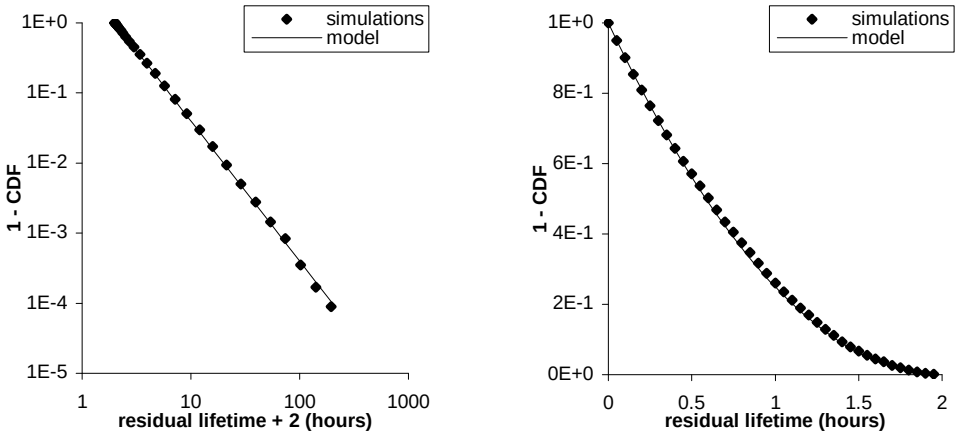
Lemma 17. *For all $j \geq 1$, the CDF of residual lifetime Z_j of the initial neighbor that starts the j-th cycle converges to* (205) *as system age approaches infinity.*

It is important to emphasize that Lemma 17 holds when switching occurs in DHTs in response to Poisson user arrivals into the system and may not hold otherwise. When a neighbor pointer switches to a new user, it loses track of which peer on the ring will be the neighbor that will start the next cycle in the link's ON/OFF process. Hence, neighbor selection during link recovery is essentially uniformly random among the existing neighbors (due to random hash indexes) and independent of the selected neighbor's age.

Given Lemma 17, the mean residual lifetime $E[Z_j]$ can be expressed directly using the properties of L as [91]:

$$E[Z_j] = \frac{E[L^2]}{2E[L]}. \tag{206}$$

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



(a) Z_2 for Pareto L with $\alpha = 3$ (b) Z_3 for uniform L

Fig. 25. Comparison of simulation results to model (205) in a deterministic DHT with $E[N] = 1,000$. In both cases, $E[L] = 1$ hour.

Simulations of Z_j for $j = 2, 3$ and two lifetime distributions are shown in Fig. 25. As demonstrated by the figure, Lemma 17 correctly predicts that recovery obtains neighbors whose residuals can be considered drawn uniformly randomly from the system and whose residual lifetimes are given by (205). This result holds for both heavy-tailed (e.g., Pareto) and light-tailed (e.g., uniform) user lifetimes. Additional simulations for larger j and other lifetime distributions confirming (205) are not shown here for brevity.

6.4.2 Exponential Lifetimes

We start by investigating $R(y)$ under exponential lifetimes. Assume that user lifetimes L are exponential with rate $\mu = 1/E[L]$. Then, it is easy to obtain from Lemma 17 that residual lifetime Z_j of the initial neighbor, for all cycles $j \geq 1$, is exponential with the same rate μ. Using $L \sim \exp(\mu)$ and $Z_j \sim \exp(\mu)$ and invoking Theorem 15 leads to the following result.

Theorem 17. *For user lifetimes L with CDF $1-e^{-\mu x}$, link lifetime R_j is independent of remaining zone size Y_j and has the same distribution as L:*

$$P(R_j < x) = 1 - e^{-\mu x}, \quad \textit{for all } j \geq 1, \tag{207}$$

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

Proof. Using the fact that neighbor residual lifetimes Z_j and user lifetimes L have the same exponential distribution with parameter $\mu = 1/E[L]$, we obtain the sojourn time τ_i in state $i \geq 0$ from (184):

$$P(\tau_i > t) = P(W_i > t)P(L > t) = e^{-(\lambda_i+\mu)t}, \tag{208}$$

where λ_i is the arrival rate given in (185). This means that τ_i is an exponential random variable with rate $\lambda_i + \mu$. Next, transition probabilities $p_{i,i+1}$, $i \geq 0$, can be computed from (186) as:

$$p_{i,i+1} = P(W_i < L) = \frac{\lambda_i}{\lambda_i + \mu}. \tag{209}$$

Then, using (208) and (209), we easily get the Laplace transform $\hat{q}_{i,i+1}(s)$ from (194):

$$\hat{q}_{i,i+1}(s) = p_{i,i+1}\frac{\lambda_i + \mu}{\lambda_i + \mu + s} = \frac{\lambda_i}{\lambda_i + \mu + s}. \tag{210}$$

Similarly, we obtain the Laplace transform $\hat{q}_{iF}(s)$:

$$\hat{q}_{iF}(s) = (1 - p_{i,i+1})\,\frac{\lambda_i + \mu}{\lambda_i + \mu + s} = \frac{\mu}{\lambda_i + \mu + s}. \tag{211}$$

Invoking Theorem 15 and substituting (210) and (211) into (196), we get the

Laplace transform of $R(y)$ for exponential lifetimes:

$$\hat{R}(s,y) = \mu\Big(\frac{1}{\lambda_0 + C} + \frac{\lambda_0}{\lambda_0 + C} \cdot \frac{1}{\lambda_1 + C} + \frac{\lambda_0}{\lambda_0 + C} \cdot \frac{\lambda_1}{\lambda_1 + C} \cdot \frac{1}{\lambda_2 + C} + \ldots\Big), \tag{212}$$

where $C = \mu + s$. Recalling that $\lambda_{i+1} = \lambda_i/2$ and setting $a = \lambda_0$, (212) reduces to:

$$\hat{R}(s,y) = \mu\Big(\frac{1}{a + C} + \frac{a}{a + C} \cdot \frac{1}{a/2 + C} + \frac{a}{a + C} \cdot \frac{a/2}{a/2 + C} \cdot \frac{1}{a/4 + C} + \ldots\Big) \doteq \mu f(C),$$

where $f(C)$ is defined as the summation term in the last equation. Observe that $f(C)$ can be transformed into:

$$f(C) = \frac{1}{a + C} + \frac{a}{a + C} \cdot \frac{2}{a + 2C} + \frac{a}{a + C} \cdot \frac{a}{a + 2C} \cdot \frac{4}{a + 4C} + \ldots = \frac{1}{a + C}\Big(1 + 2a \cdot f(2C)\Big). \tag{213}$$

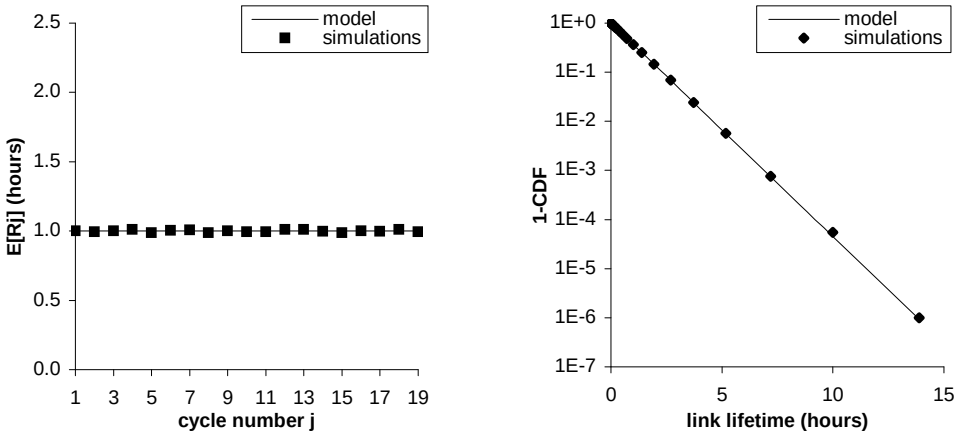
Solving the last recurrence, we have $f(C) = 1/C$, which is the only solution since the infinite summation $f(C)$ is a unique real number (convergence follows from the monotonically increasing nature of the summation as a function of the number of terms). We finally obtain:

$$\hat{R}(s,y) = \frac{\mu}{C} = \frac{\mu}{\mu + s}, \tag{214}$$

which shows that $R(y)$ is an exponential variable with parameter μ. It is apparent from (214) that $R(y)$ is independent of $Y_j = y$, which then establishes this theorem.

□

Model (207) is very accurate as shown in Fig. 26. Notice from the left figure that $E[R_j]$ is equal to mean user lifetime $E[L]$ and from the right figure that the



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

Fig. 26. Comparison of model (207) to simulations in a deterministic DHT with $E[N] = 2,000$ and exponential user lifetimes with $E[L] = 1$ hour.

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

The rationale behind Theorem 17 can be explained as follows. Recall that Z_j is the residual lifetime of the first neighbor u that owns the neighbor pointer in each cycle. Due to the memoryless property of exponential distributions, the remaining time of Z_j obtained at a random instant is still exponential with rate μ, which matches the lifetime distribution of new arrivals entering the same zone. Therefore, it makes no difference whether a current neighbor u is replaced by a new arrival or not. Then, it is not hard to see that the link lifetime has the same distribution as Z_j, which is $\exp(\mu)$. A similar scenario is observed in $M/M/1$ queues [91] where customers can be interrupted during services and the distribution of the total service time required for a customer does not change.

Theorem 17 indicates that switching has no impact on link lifetimes in any DHT with exponential user lifetimes, which makes analysis of system performance in such systems very simple. However, we should note that this result does not hold for any

non-exponential lifetime distribution. As recent measurements of P2P networks show that user lifetimes are often heavy-tailed [12], [89], we next use the Pareto distribution $P(L < x) = 1 - (1 + x/\beta)^{-\alpha}$ with shape parameter $\alpha > 1$ and scale parameter $\beta > 0$ to estimate the performance of real DHTs under churn.

6.4.3 Pareto Lifetimes

For Pareto L, it is clear from Lemma 17 that the residual lifetime Z_j of initial neighbors follows the CDF $P(Z_j < x) = 1-(1+x/\beta)^{-(\alpha-1)}$ for all $j \geq 1$, which shows that Z_j are also Pareto-distributed but more heavy-tailed. Next, we apply Theorem 15 to obtain the Laplace transform $\hat{R}(y, s)$ and Theorem 16 to obtain the mean of $R(y)$.

Theorem 18. *For Pareto lifetimes L, the mean conditional link lifetime $E[R(y)]$ is given by* (202) *with*

$$E[\tau_i] = \beta e^{\lambda_i \beta} E_{\alpha_i}(\lambda_i \beta), \quad p_{i,i+1} = \lambda_i E[\tau_i] \tag{215}$$

where arrival rate λ_i is given in (185)*, $E_k(x) = \int_1^\infty e^{-xu}\ u^{-k} du$ is the generalized exponential integral, and*

$$\alpha_i = \begin{cases} \alpha - 1 & i = 0 \\ \alpha & i \geq 1 \end{cases}. \tag{216}$$

Furthermore, the Laplace transform $\hat{R}(y, s)$ is given by (196) *with*

$$\hat{q}_{i,i+1}(s) = \lambda_i E[\tau_i] A, \quad \hat{q}_{iF}(s) = (1 - \lambda_i E[\tau_i]) A, \tag{217}$$

where $A = 1 + (1 - \lambda_i - s)\beta e^{(\lambda_i+s)\beta} E_{\alpha_i}((\lambda_i + s)\beta)$, and $E[\tau_i]$ is shown in (215) *and α_i in* (216)*.*

Proof. Since $Z_j \sim \text{Pareto}(\alpha - 1, \beta)$ for all $j \geq 1$, we obtain the distribution of sojourn

time τ_0 in state 0 from (184):

$$\begin{aligned} P(\tau_0 > t) &= P(W_0 > t)P(Z_j > t) \\ &= e^{-\lambda_0 t}\Big(1 + \frac{t}{\beta}\Big)^{-(\alpha-1)}, \end{aligned} \tag{218}$$

where λ_0 is given in (185). Then, we easily get the PDF of τ_0:

$$\begin{aligned} f_{\tau_0}(t) = -\frac{dP(\tau_0 > t)}{dt} &= \lambda_0 e^{-\lambda_0 t}\Big(1 + \frac{t}{\beta}\Big)^{-(\alpha-1)} \\ &+ \frac{\alpha - 1}{\beta} e^{-\lambda_0 t}\Big(1 + \frac{t}{\beta}\Big)^{-\alpha}, \end{aligned} \tag{219}$$

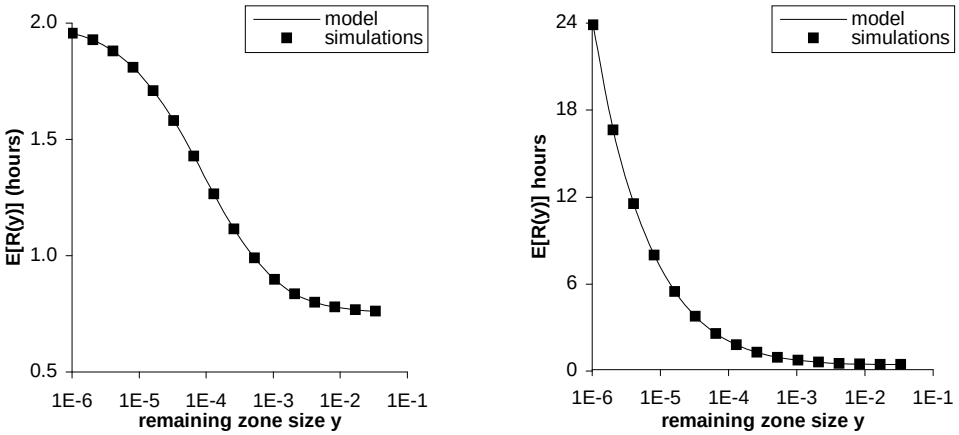
and its mean:

$$\begin{aligned} E[\tau_0] = \int_0^\infty P(\tau_0 > t)dt &= \int_0^\infty e^{-\lambda_0 t}\big(1 + \frac{t}{\beta}\big)^{-\alpha+1} dt \\ &= \beta e^{\lambda_0 \beta} E_{\alpha-1}(\lambda_0 \beta), \end{aligned} \tag{220}$$

where $E_k(x) = \int_1^\infty e^{-xu}\ u^{-k} du$ is the generalized exponential integral. Next, the transition probability p_{01} from state 0 to 1 can be computed from (186) as:

$$\begin{aligned} p_{01} = P(W_0 < Z_j) &= \int_0^\infty P(W_0 < t) f_Z(t) dt \\ &= \int_0^\infty \left(1 - e^{-\lambda_0 t}\right) \frac{\alpha - 1}{\beta} \left(1 + \frac{t}{\beta}\right)^{-\alpha} dt \\ &= 1 - (\alpha - 1) e^{\lambda_0 \beta} E_\alpha(\lambda_0 \beta) \\ &= \lambda_0 \beta e^{\lambda_0 \beta} E_{\alpha-1}(\lambda_0 \beta) = \lambda_0 E[\tau_0], \end{aligned} \tag{221}$$

where the last step is established upon recalling (220). Substituting (219) and (221) into (194) and doing certain algebra, we obtain the Laplace transforms of the semi-



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

Fig. 27. Comparison of model $E[R(y)]$ in Theorem 18 to simulation results in a deterministic DHT with mean size $E[N] = 2,000$ and Pareto user lifetimes L with mean $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.

Markov kernel starting from state 0:

$$\hat{q}_{01}(s) = p_{01} \int_0^\infty e^{-st} f_{\tau_0}(t)dt = \lambda_0 E[\tau_0] \times [1 + (1 - \lambda_0 - s)\beta e^{(\lambda_0+s)\beta} E_{\alpha-1}((\lambda_0 + s)\beta)], \quad (222)$$

$$\hat{q}_{0F}(s) = (1 - \lambda_0 E[\tau_0]) \left[1 + (1 - \lambda_0 - s)\beta \times e^{(\lambda_0+s)\beta} E_{\alpha-1}((\lambda_0 + s)\beta)\right]. \quad (223)$$

Laplace transforms $\hat{q}_{i,i+1}(s)$ and $\hat{q}_{iF}(s)$, $i \geq 1$ can be obtained by replacing λ_0 with λ_i and $\alpha - 1$ with α in the above equations. Invoking Theorems 15-16, we have the desired result. □

Fig. 27 shows simulation results of $E[R(y)]$ for several values of remaining zone sizes y and the plots the corresponding model from Theorem 18. Besides the accuracy of the model, notice from this figure that as remaining zone size y reduces, $E[R(y)]$ increases and converges to $E[Z_1]$, where the distribution of neighbor residual lifetime

Z_1 is given in (205).

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

6.4.4 Zone Sizes

In order to determine the distribution of zone sizes U and Y_j in Fig. 23, we must decide on the zone splitting method. The derivations below only cover the random-split [90] mechanism (i.e., zones are split at hash indexes of arriving users) that is used in Chord [79] and only considers one-dimensional DHTs. A similar derivation can be carried out for the center-split [52], [67] strategy (i.e., zones are always split in the center) and multi-dimensional DHTs, but this analysis is much more tedious and is not shown here.

Since all arriving users are placed in the interval $[0, 1)$, the average zone size is approximately $1/E[N]$, where N is the random system size in the steady-state. Approximation $E[1/N] = 1/E[N]$ is asymptotically accurate as system size tends to infinity for the ON/OFF churn model of [93]. This follows from the fact that $N/E[N]$ converges to 1 in probability. The next result states that in equilibrium DHTs, zone sizes no larger than $1/\sqrt{E[N]}$ are distributed approximately exponentially. Since most zone sizes do not deviate from the mean very far, this result directly applies to random variable U defined earlier.

Lemma 18. *As the mean system size tends to infinity, the distribution of small zones in the DHT becomes approximately exponential:*

$$\lim_{E[N]\to\infty} \frac{P(U > x)}{e^{-E[N]x}} = 1 \tag{224}$$

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

Proof. We assume that the probability that a user of any given zone size departs is equally likely (i.e., zone sizes do not depend on user lifetimes and vice versa). Then, given that hash index X_i of any user i is uniformly random in $[0, 1)$ at any time t, it is well-known that zone sizes U are uniformly distributed on the simplex $\{(x_1, \cdots, x_N)|x_i \geq 0; \sum x_i = 1\}$ [17]. It follows that conditioning on $N = z$, the probability that a zone of size x from a given point X_i of user i is unoccupied by the remaining $z - 1$ users is simply:

$$P(U > x|N = z) = (1 - x)^{z-1}. \tag{225}$$

Note that $(1 - x)^{z-1}$ can be transformed into:

$$(1 - x)^{z-1} = e^{(z-1)\log(1-x)} = e^{-x(z-1)+O(x^2)(z-1)}, \tag{226}$$

where the expansion uses the Taylor approximation of $\log(1 - x)$. Substituting (226) into (225) and keeping in mind that $x = o(1/\sqrt{E[N]})$, we obtain:

$$\frac{P(U > x|N = z)}{e^{-xz}} = e^{x+O(x^2)(z-1)} \to 1, \tag{227}$$

as $E[N] \to \infty$.

For the heterogeneous user churn model, recall from [93, Lemma 1] that N is a Gaussian variable with PDF $f_N(z)$. The distribution $P(U > x)$ can then be computed by integrating $P(U > x|N = z)$ with respect to z:

$$\lim_{E[N]\to\infty} \frac{P(U > x)}{e^{-E[N]x}} = \frac{\int_0^\infty e^{-xz} f_N(z)dz}{e^{-E[N]x}}, \tag{228}$$

where the last step is obtained by using (227). It then follows from (228) that:

$$\lim_{E[N]\to\infty} \frac{P(U > x)}{e^{-E[N]x}} = \frac{e^{-E[N]x+Var[N]x^2/2}}{e^{-E[N]x}}, \tag{229}$$

since e^{-xN} is a lognormal random variable. Recalling $Var[N] < E[N]$ [93, Lemma 1]

and $x^2 E[N] \to 0$ as $E[N] \to \infty$, (229) yields:

$$\lim_{E[N]\to\infty} \frac{P(U > x)}{e^{-E[N]x}} = 1, \tag{230}$$

which is the desired result. Finally, note that the requirement of $x^2 E[N] \to 0$ is tight and cannot be relaxed for computing the distribution of U. □

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

Lemma 19. *For a given zone size y, assume that $y^2 E[N] \to 0$ as $E[N] \to \infty$. Then, the PDF $f_{Y_j}(y)$ of remaining zone size Y_j is asymptotically:*

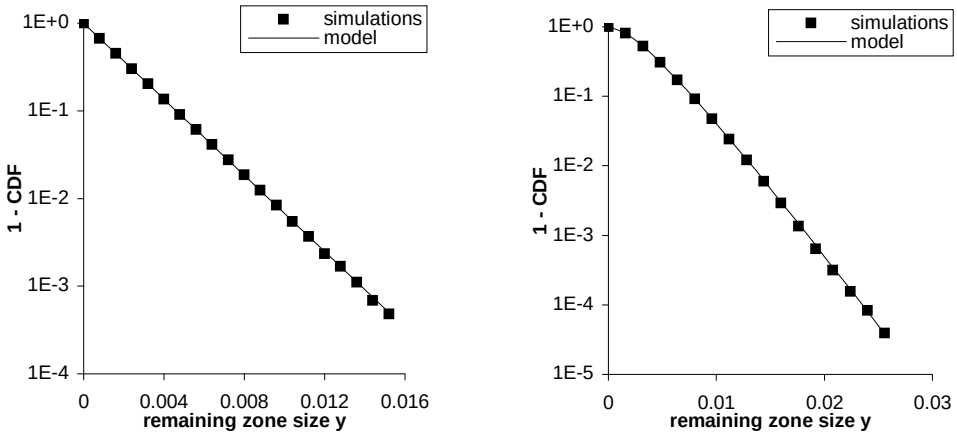
$$\begin{cases} \lim\limits_{E[N]\to\infty} \dfrac{f_{Y_1}(y)}{E[N]e^{-E[N]y}} = 1 & j = 1 \\ \lim\limits_{E[N]\to\infty} \dfrac{f_{Y_j}(y)}{E[N]^2 y e^{-E[N]y}} = 1 & j \geq 2 \end{cases}, \tag{231}$$

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

Proof. Due to the memoryless property of the exponential limiting distribution of U shown in (224), the remaining zone size Y_1 from a neighbor pointer, which randomly splits the zone of some neighbor u, to the hash index of u follows the same distribution of U.

Next, note that Y_j, $j \geq 2$, is the initial zone size of a replacement neighbor u obtained by user v during each recovery. At this time, replacement neighbor u covers its own zone as well as that of the failed user. Thus, it is clear that $Y_j = Y_1 + U$, which has the same distribution as $U + U$. It then immediately follows that Y_j, $j \geq 2$, has the Erlang-2 distribution since it is a sum of two exponentials. □

Lemma 19 shows that the distribution of Y_1 is exponential and that of Y_j for $j \geq 2$ is Erlang-2. As demonstrated in Fig. 28, model (231) is very accurate even for



(a) distribution of Y_1 (b) distribution of Y_2

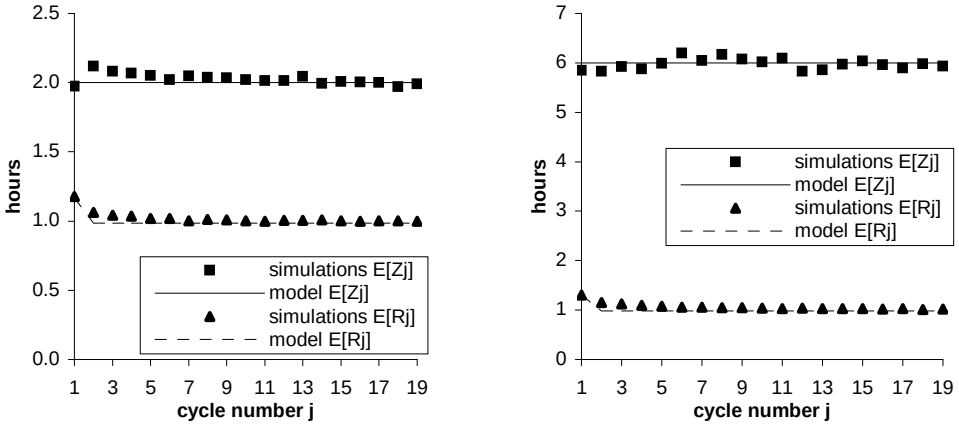
Fig. 28. Comparison of simulation results of Y_j to model (231) in a deterministic DHT with mean size $E[N] = 500$ under churn produced by Pareto L with $\alpha = 3$ and $E[L] = 1$ hour.

small average system size $E[N] = 500$ users. Additional simulation results confirming (231) for larger $E[N]$ and different j are not shown for brevity.

6.4.5 Putting the Pieces Together

The final step is to apply (181) and (182) to uncondition the distribution of link lifetime R_j and its mean $E[R_j]$ using the distribution of initial zone size Y_j given in (231). To this end, substituting $E[R(y)]$ shown in Theorem 18 and the PDF of Y_j in (231) into (182) leads to the final result on the mean link lifetime $E[R_j]$. Similarly, to get the distribution of R_j, we first retrieve the distribution of $R(y)$ from $\hat{R}(s, y)$ in Theorem 18 by applying an existing inverse Laplace transform software package [1]. Then substituting the distribution of $R(y)$ and (231) into (181) leads to the final model of the distribution of link lifetime R_j.

Fig. 29 shows simulations results and the model of the mean link lifetime $E[R_j]$ and the average residual lifetime $E[Z_j]$ of the initial neighbor that starts the j-th cycle.

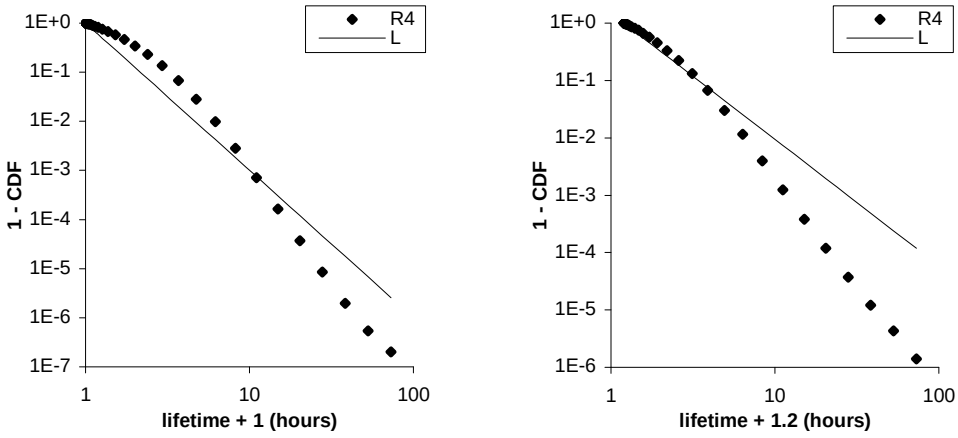


(a) $\alpha = 3$ (b) $\alpha = 2.2$

Fig. 29. Comparison of $E[R_j]$ to $E[Z_j]$ in a deterministic DHT with mean size $E[N] = 2,500$ users, Pareto lifetimes with mean $E[L] = 1$ hour, and $\beta = E[L](\alpha - 1)$.

The model of $E[Z_j]$ is obtained using (206) and the general solution to $E[R_j]$ is given in (182). As shown in the figure, both models match simulation results very well and as α becomes smaller, the difference between $E[R_j]$ and $E[Z_j]$ increases as expected. Recall that smaller α leads to stochastically larger Z_j and thus increases reliability of non-switching systems [42]. The above results also show that the process of switching to new users can significantly reduce the lifetime of a link and that deterministic DHT systems with Pareto L can exhibit $E[R_j]$ very close to $E[L]$. This is in contrast to unstructured P2P systems where $E[R_j]$ can be $11 - 16$ times higher than $E[L]$ depending on shape parameter α [12], [89].

Further observe from the model and Fig. 29 that link lifetimes are completely characterized by two random variables R_1 and R_2 since R_j for $j \geq 3$ has the same distribution as R_2. This arises from the fact that zone size Y_1 is different from Y_2, while Y_j for $j \geq 3$ are all distributed as Y_2. Since Y_1 is stochastically smaller than Y_2 (see Lemma 19), it follows that R_1 is stochastically larger than R_2. Furthermore,



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

Fig. 30. Link lifetimes R_4 are less heavy-tailed than Pareto user lifetimes L in a deterministic DHT with mean size $E[N] = 2,500$ peers, $E[L] = 1$ hour, and $\beta = (\alpha - 1)E[L]$.

from the analysis of the Markov chain in previous sections, it becomes clear that selecting neighbors with *smaller* initial zone sizes leads to larger link lifetimes since such neighbors are less likely to be replaced by newly arriving users and the link's $E[R_j]$ will be closer to $E[Z_j]$.

The most intriguing result shown in Fig. 29 is that $E[R_j]$ for all $j \geq 2$ is very close to the mean user lifetime $E[L]$ under different values of α (e.g., $E[R_4] = 0.986$ hours for $\alpha = 3$ and 1.096 for $\alpha = 2.2$). However, from the model of the tail distribution of link lifetime R_4 shown in Fig. 30, observe that the distribution of R_j for $j \geq 2$ is actually different from that of lifetime L and is *less* heavy-tailed than the original distribution. A similar result holds for other values of α and other distributions, which we do not show for brevity.

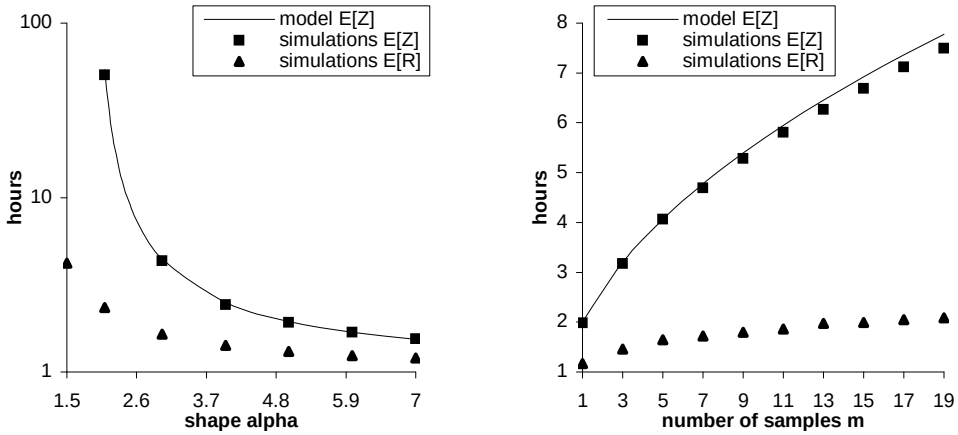
6.5. Randomized DHTs

Since the user arrival process into a DHT usually cannot be changed to achieve better system resilience, peers may utilize the knowledge of residual lifetime Z_j of the initial owner of a given link and/or remaining zone size Y_j to improve link lifetime R_j. In the following, we make use of the freedom of selecting links in randomized DHTs to achieve the goal of increasing R_j using two different link-selection strategies.

6.5.1 Max-Age Selection

The first strategy we apply for selecting neighbor pointers is called *max-age* [84], [96]. In this technique, which we explain using the example of Randomized Chord [29], user v with hash index $id(v) \in [0, 1)$ uniformly randomly samples m points in the range $[id(v) + 2^i/2^{64}, id(v) + 2^{i+1}/2^{64})$ and selects the point whose successor has the maximum age as its i-th neighbor pointer. Note that switching occurs as described before (i.e., when new users split a given zone and replace existing neighbors) and link failure is repaired by replacing the dead neighbor (i.e., the last user holding the link) with the current successor.

It is clear that link lifetimes R_j for all cycles $j \geq 1$ have the same distribution since the neighbor pointer in each cycle is uniformly randomly generated within a certain range of users (as mentioned before, we assume the range is large enough to support non-trivial choices). Simulation results of max-age selection and the model of $E[Z_j]$ from [96] are shown in Fig. 31. First notice from part (a) that for a fixed number of samples $m = 6$, as shape α decreases, the mean link lifetime $E[R_j]$ increases much slower than the mean residual lifetime $E[Z_j]$ of the initial neighbor (in fact, $E[Z_j] = \infty$ for $\alpha \leq 2$). A similar phenomenon appears in part (b) where $E[Z_j]$ increases at the rate of $\sqrt{m}$ for $\alpha = 3$ (see [96, Lemma 5]), while $E[R_j]$ rises



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

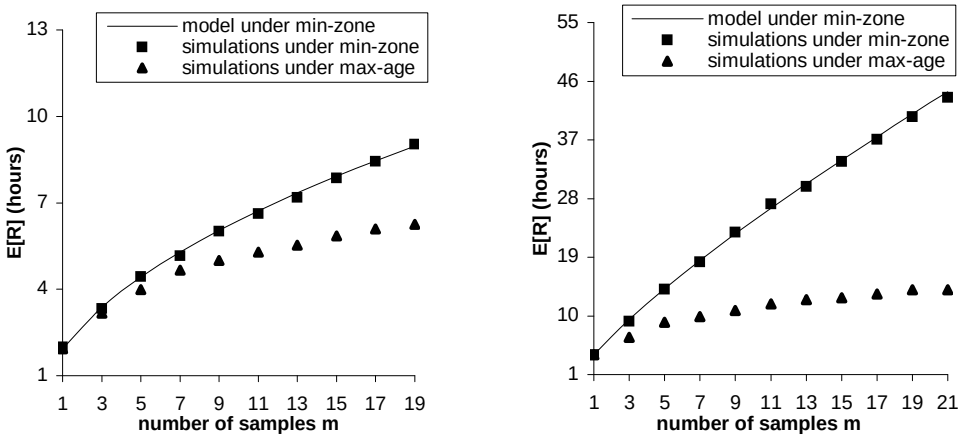
Fig. 31. Impact of shape α and number of samples m on mean link lifetime $E[R_j]$ under max-age selection in a randomized DHT with mean size $E[N] = 2,000$ for Pareto lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.

from 1.17 hours to only 2.09 hours as m increases from 1 to 19. These two subfigures demonstrate that the improvement in terms of the mean link lifetime $E[R_j]$ under max-age selection is generally very small since new arrivals sooner or later split initial neighbors to take ownership of the link and hence ages or residual lifetimes of original neighbors do not affect link churn rate very much.

6.5.2 Min-Zone Selection

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

To obtain a model for $E[R_j]$ under min-zone selection, first note that residual lifetime Z_j of the initial neighbor starting the j-th cycle follows the distribution given in (205) since all m samples are uniformly random and zone sizes are independent



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

Fig. 32. Comparison of mean link lifetime $E[R_j]$ under min-zone selection to that under max-age selection in a randomized DHT with mean size $E[N] = 2,000$ for Pareto user lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.

of user ages or lifetimes. It is then clear that for a fixed remaining size $Y_j = y$, the Laplace transform and the mean conditional link lifetime given in Theorem 18 are both still valid. Next, given that initial zone size Y_j is minimum among m uniformly randomly selected samples, we readily obtain:

$$P(Y_j > y) = [P(U > y)]^m, \quad \text{for all } j \geq 1, \tag{232}$$

where U is the zone size of a randomly selected user on the ring whose limiting distribution is shown in (224). The final step is to combine Theorem 18 and (232) to obtain the distribution of R_j and its mean under min-zone selection.

As shown in Fig. 32, the model of $E[R_j]$ matches simulation results very well. Most interestingly, the figure demonstrates that the mean link lifetime $E[R_j]$ under min-zone selection is significantly larger than that under max-age selection for both choices of α and that the difference between the two metrics becomes more pronounced as the number of samples m increases or shape α decreases. Furthermore, this figure

suggests that as $m \to \infty$, $E[R_j]$ for min-zone selection and $\alpha < 2$ goes to infinity, while $E[R_j]$ for max-age selection converges to some fixed number regardless of α. The following theorem confirms this result.

Theorem 19. *For Pareto user lifetimes with $1 < \alpha \leq 2$, the expected link lifetime under min-zone selection approaches infinity for sufficiently large system population and random sample size:*

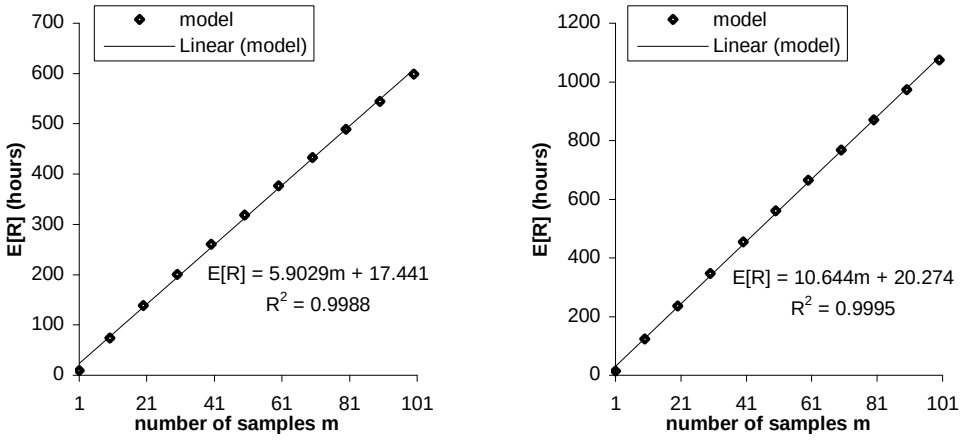
$$\lim_{E[N]\to\infty} \lim_{m\to\infty} E[R_j] = \infty. \tag{233}$$

For max-age selection and any α, the mean link lifetime converges to a constant:

$$\lim_{E[N]\to\infty} \lim_{m\to\infty} E[R_j] < \infty. \tag{234}$$

Proof. To obtain $E[R_j]$ under min-zone selection for $m \to \infty$, first note from (232) that $P(Y_j > y) \to 0$ as $m \to \infty$ for all fixed $y > 0$. This indicates that $Y_j \to 0$ in probability. It is then clear that the probability that a new arrival splits a given zone with size Y_j also approaches 0, and hence in the limit R_j is simply residual lifetime Z_j of the initial neighbor. Recalling from (205) that $E[Z_j] = \infty$ for $\alpha \leq 2$, we immediately obtain $E[R_j] \to E[Z_j] = \infty$ as $m \to \infty$. The condition $E[N] \to \infty$ is required for $m \to \infty$.

When max-age selection is used, it is shown in [96, Theorem 5] that residual lifetimes $Z_j \to \infty$ with probability 1 as $m \to \infty$ for Pareto lifetimes. It is then easy to obtain using the semi-Markov chain $\{A_\delta^y\}$ in Theorem 1 that sojourn time τ_0 in state 0 is $\min(Z_j, W_0) \to W_0$ as $m \to \infty$, where W_0 is exponential with rate λ_0 given in (185), and transition probability $p_{0,1} = P(W_0 < Z_j) \to 1$. After the chain jumps into state 1, sojourn times are $\min(L, W_i)$, which are no longer affected by the number of samples m. Hence, $E[R_j]$ is finite since the mean sojourn time in each state i is finite and the probability that the chain jumps into the failed state increases



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

Fig. 33. Approximation of $E[R_j]$ as a linear function of number of samples m under min-zone selection for Pareto user lifetimes with $E[L] = 1$ hour and $\beta = E[L](\alpha - 1)$.

exponentially fast. □

The above analysis indicates that min-zone selection is significantly better than max-age selection for very heavy-tailed user lifetimes. Since real systems have been observed to exhibit $\alpha \approx 1.06$ in [12] and $\alpha = 1.09$ in [89], this result paves a simple way for building better DHTs in practice. The amount of actual improvement in $E[R_j]$ for these two values of α is shown in Fig. 33, where the growth rate in both curves is approximately linear in m. The figures also show the corresponding linear fits to the model, which can be used to predict how m affects link lifetime $E[R_j]$ in these two cases. For instance, with $\alpha = 1.09$, users can obtain $E[R_j] \approx 76$ hours by sampling $m = 10$ points for each suitable (i.e., with enough random choices) link in a randomized DHT. For $\alpha = 1.06$, the corresponding average link lifetime is 127 hours. Comparing these numbers to $E[R_j] \approx E[L] = 1$ hour in deterministic DHTs, the extent of improvement is undoubtedly dramatic.

6.6. Summary

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

CHAPTER VII

SUCCESSOR LISTS IN DHTS

7.1. 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 [2], [6], [16], [29], [34], [40], [42], [50], [61], [71], [80]. While previous analytical work [42], [45] 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* [72], [80], whose sole purpose is to recover the system from inconsistent states and provide resilience [80]. In this chapter, we focus on partitioning of one particular Distributed Hash Table (DHT) called Chord [80] and note that similar results can be obtained for other types of successor/leaf sets.

Recall that each node v in Chord maintains a list consisting of its $r = \Theta(\log n)$ successors and a routing table containing $k = \Theta(\log n)$ neighbor pointers, where n is the system size. Note that routing tables are used to reduce lookup latency, while successors ensure resilience during churn. 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. When all r successors of any node fail simultaneously, the system becomes partitioned and is potentially unable to recover without a bootstrap. Although neighbors in some routing tables may still be alive, there is no guarantee that the system can return to a consistent state after partition-

ing. 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:

$$P(\text{graph disconnects}) = P(X > 0), \tag{235}$$

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 chapter, previous neighbor churn models [42] cannot be applied to obtain the probability in (235). We perform this task below for both static and dynamic node failure.

7.1.1 Static Failure

Many prior studies have been interested in the resilience of structured P2P networks against *static* node failure [29], [34], [80], i.e., when each node independently fails with a certain probability p. We apply the Erdös-Rényi theorem to show that under p-fraction node failure, the probability that Chord with size $n \to \infty$ remains connected is asymptotically:

$$\lim_{n\to\infty} \frac{P(X=0)}{e^{-n(1-p)p^r}} = 1, \tag{236}$$

where $r = \Theta(\log n)$ is the number of immediate successors a user monitors. It is rather surprising to find from (236) 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 [45]. Setting $r = c \log_2 n$, where $c > 0$ is a constant, (236) shows that as $n \to \infty$ the probability that Chord remains connected approaches 1 if $p < 2^{-1/c}$ and 0 if $p > 2^{-1/c}$.

7.1.2 Dynamic Failure

As observed in deployed structured P2P file-sharing systems [64], [81], users join and fail at a high rate of churn. The second part of this chapter 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 $F(x)$. User v then 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:

$$\phi \leq \frac{\rho\rho!r!}{(\rho+r)!}, \tag{237}$$

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

We then examine how individual node isolation affect partitioning of the system as nodes continuously join and leave. Using the Chen-Stein method [5], we establish that when $r \to \infty$ the probability that Chord stays connected after experiencing N user joins is asymptotically:

$$\lim_{N\to\infty} \frac{P(X=0)}{(1-\phi)^N} = 1, \tag{238}$$

where ϕ is the node isolation probability given in (237). 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 [45] without proof for independent neighbor behavior in routing tables, our result in (238) is again for dependent node isolations and is formally proven.

As (238) 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 [80]. Our results show that both stabilization strategies are much better than the exponential S suggested in [39], 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:

$$\phi \approx \frac{\rho\rho!}{(\rho + r)!}, \tag{239}$$

where $\rho = E[L]/E[S]$. The amount of improvement over the exponential version (237) of this metric is by a factor of $r!$, which is significant in most cases.

We finish the chapter by studying non-exponential lifetimes observed in real P2P graphs [89]. Even though models of ϕ for heavy-tailed user lifetimes are currently intractable, we show that ϕ in such systems is upper bounded by the exponential metric (237). We confirm this effect and demonstrate the distance to the upper bound in simulations.

7.2. 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 non-human 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.

7.2.1 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 $0 \leq p \leq 1$ is a given number [29], [34], [45], [80]. 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:

$$X_i = \begin{cases} 1 & \text{user } i \text{ is alive and its } r \text{ successors failed} \\ 0 & \text{otherwise} \end{cases}. \tag{240}$$

Note that unlike [45], 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:

$$X = \sum_{i=1}^{n} X_i, \tag{241}$$

where n is the number of nodes in Chord. It is then clear from (235) 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 [45] for two reasons: 1) the model in [45] 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 [45] without a formal proof due to a much wider variety of neighbor sets covered by [45].

Theorem 20. *The probability that each user in Chord remains connected to at least one successor under p-fraction node failure is asymptotically:*

$$\lim_{n\to\infty} \frac{P(X=0)}{e^{-n(1-p)p^r}} = 1, \tag{242}$$

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:

$$p = P(Y_i = 1) = 1 - P(Y_i = 0). \tag{243}$$

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

$$L_n = \max_{1\le i\le n-k+1} \{k : Y_i = Y_{i+1} = \cdots = Y_{i+k-1} = 1\}. \tag{244}$$

Now notice 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:

$$P(X = 0) = P(L_n < r). \tag{245}$$

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 [7]:

$$\frac{P(L_n < r)}{e^{-n(1-p)p^r}} \to 1, \tag{246}$$

as $n \to \infty$, which immediately leads to (242). □

The asymptotic result in (242) allows us to utilize a very accurate approximation:

$$P(\text{Chord is connected}) = P(X = 0) \approx e^{-n(1-p)p^r}, \tag{247}$$

which we verify next in finite-size graphs. Simulation results of $P(X = 0)$ in Chord under static node failure are presented in Table IV. 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 IV 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 (247) and simulation results reduces fast. The rest of the table shows additional examples of model's accuracy for several choices of p and r.

7.2.2 Discussion

We next relate our results in Theorem 20 to those in [45, Proposition 3]. Recall that [45] defines isolation as an event of a user losing all of its neighbors in Fig. 21(b). Their results show that all users have at least one alive neighbor with probability:

$$P(X = 0) \approx e^{-n(1-p)p^k}, \tag{248}$$

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 (247) 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 [45] (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

Table IV. Comparison of simulation results of $P(X = 0)$ under static node failure to model (247) 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	(247)	p	Simulations	(247)	p	Simulations	(247)	p	Simulations	(247)
$1,000$	.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

its r successors is simply:

$$\phi = P(X_i = 1) = (1-p)p^r, \quad 1 \leq i \leq n \tag{249}$$

where X_i is the Bernoulli variable defined in (240). Note that given that $r \to \infty$ as $n \to \infty$, it is readily seen from (249) that $\phi \to 0$ as $n \to \infty$. Using (249), the approximation in (247) can be transformed into:

$$P(X = 0) \approx e^{-n\phi} \approx (1-\phi)^n, \tag{250}$$

where Taylor expansion $e^{-x} = 1 - x$ holds for small enough x as $n \to \infty$. Thus, (250) indicates that

$$P(X = 0) = P\Big(\bigcap_{i=1}^{n}[X_i = 0]\Big) \approx \prod_{i=1}^{n} P(X_i = 0) \tag{251}$$

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 (251) can be explained by the Chen-Stein theorem [5], 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 [45]. However, the rate of convergence of $P(X = 0)$ in (247) and (248) is different.

7.3. Dynamic Node Failure: General Results

Recent measurements of P2P networks [12], [64], [81] 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 v's fail within its lifetime, and then derive the probability that Chord remains connected when stabilization intervals are exponentially distributed. We leave derivations for non-exponential intervals for the next section.

7.3.1 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) [50], 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. 34 illustrates the evolution of user v's successor list in our simple model. As

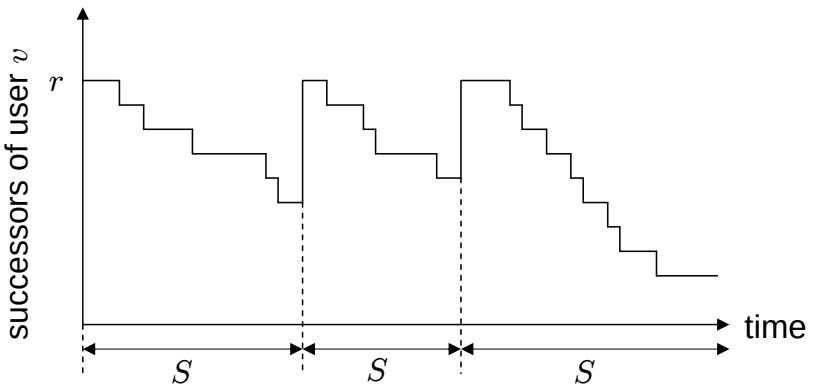


Fig. 34. Evolution of a node's successor list over time.

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 remaining lifetimes of existing successors at the start of each stabilization interval. For exponential user lifetimes, 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 that verify our theoretical results. In simulations, user arrivals occur according to a Poisson process derived in [93] for the heterogenous churn model proposed therein. 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 $F(x)$.

7.3.2 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 $Z(t) \leq r$ 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. 34 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 20. *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 $Q = (Q_{jj'})$:*

$$Q_{jj'} = \begin{cases} \theta & j \neq r, j' = r \\ j\mu & 1 \leq j \leq r, j' = j - 1 \\ -\theta - j\mu & j' = j < r \\ -j\mu & j = j' = r \\ 0 & \text{otherwise} \end{cases}, \tag{252}$$

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

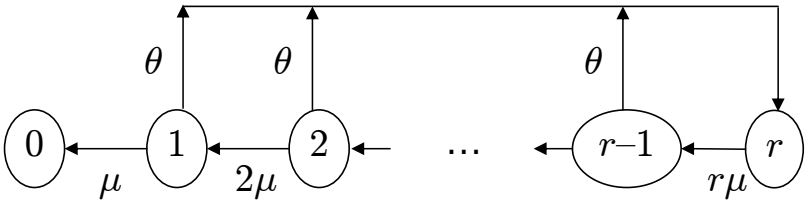


Fig. 35. Markov chain $\{Z(t)\}$ modeling a node's successor list.

Proof. We first consider state $Z(t) = r$, i.e., the full list of successors at time t (see Fig. 35). 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 $1 \leq j < r$, and the transition rate from state j to r is θ for $1 \leq j < r$. This directly leads to the desired result. □

The state diagram and transition rates of process $\{Z(t)\}$ are illustrated in Fig. 35, where each state models the number of alive successors and absorbing state 0 corresponds to user isolation. We usually write matrix Q in (252) in the canonical form:

$$Q = \begin{pmatrix} 0 & 0 \\ \mathbf{r} & Q_0 \end{pmatrix}, \tag{253}$$

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:

$$T = \inf(t > 0 : Z(t) = 0 | Z(0) = r\}. \tag{254}$$

Using Theorem 4 in Chapter IV, the isolation probability $\phi = P(T < L)$ can be reduced to:

$$\phi = \pi(0) V B V^{-1} \mathbf{r}, \tag{255}$$

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

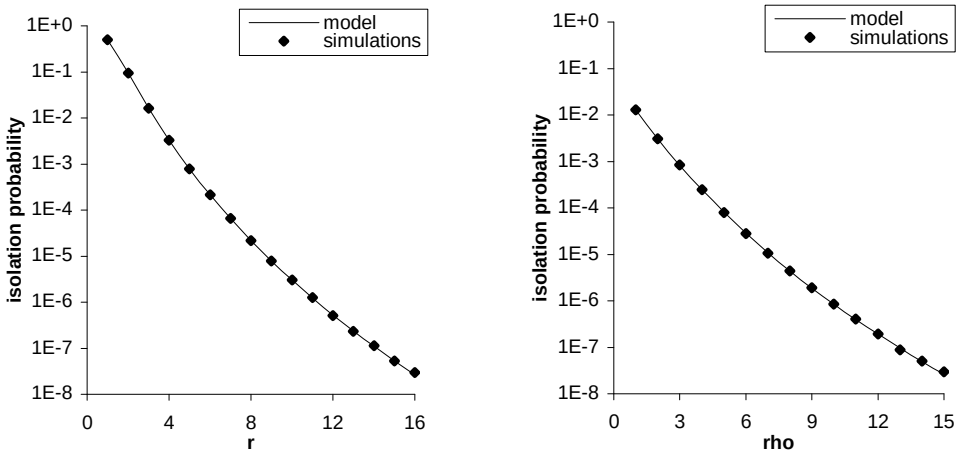
$$b_j = 1/(\mu - \xi_j), \tag{256}$$

$\mu = 1/E[L]$, $\xi_j \leq 0$ is the j-th eigenvalue of Q_0, and Q_0 and $\mathbf{r}$ are in (253).

Simulation results of isolation probability ϕ are shown in Fig. 36. Notice from this figure that model (255) is very accurate compared to simulations. Also observe that as ρ or r increase, node isolation probability sharply decreases. While (255) 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 chapter) if an explicit model of ϕ is not derived. We perform this task next.

7.3.3 Closed-Form Bounds on ϕ

Note from Fig. 34 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



(a) fixed $\rho = 15$ (b) fixed $r = 16$

Fig. 36. Comparison of model (255) to simulation results on node isolation probability ϕ for exponential lifetimes with $E[L] = 0.5$ hours and exponential stabilization intervals with $E[S] = E[L]/\rho$.

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:

$$f = P(\max\{L_1, \ldots, L_r\} < S), \tag{257}$$

where $L_i \sim \exp(\mu)$ is the remaining lifetime of the i-th successor at the beginning of a particular interval. Then, from Jensen's inequality [38, 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 $E[L]/E[S] \to \infty$.

Theorem 21. *For $L \sim \exp(\mu)$ and $S \sim \exp(\theta)$, isolation probability ϕ is upper-bounded by:*

$$\phi < \rho f, \tag{258}$$

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:

$$\lim_{\rho\to\infty} \frac{\phi}{\rho f} = 1. \tag{259}$$

Proof. Given that $S \sim \exp(\theta)$, probability f that all r successors fail with a particular interval S in (257) reduces to:

$$f = \int_0^\infty \left(1 - e^{-\mu t}\right)^r \theta e^{-\theta t} dt. \tag{260}$$

Setting $\rho = \theta/\mu$ and $z = 1 - e^{-\mu t}$, (268) yields:

$$f = \rho\mu \int_0^1 z^r (1-z)^\rho \frac{1}{\mu(1-z)} dz = \frac{\rho! r!}{(\rho + r)!}. \tag{261}$$

It is ready to see from (261) that as $\rho \to \infty$ and/or $r \to \infty$, $f \to 0$.

Next, note from Fig. 34 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:

$$\phi(j) = 1 - (1 - f)^j, \quad j \geq 1, \tag{262}$$

where $(1 - f)^j$ is the probability that user v survives all j stabilization intervals and f is given in (261).

It is then clear from Jensen's inequality [38] in the discrete form that for concave function $\phi(j)$ shown in (262), the unconditional isolation probability ϕ yields:

$$\phi = E[\phi(D)] \leq 1 - (1 - f)^{E[D]}, \tag{263}$$

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:

$$E[D(t)] = \theta t, \quad \text{for all } t \geq 0. \tag{264}$$

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

$$E[D] = \int_0^\infty E[D(t)] f_L(t) dt, \tag{265}$$

where $f_L(t)$ is the PDF of user lifetimes L. Substituting (264) into the above readily leads to:

$$E[D] = \theta E[L] = \rho, \tag{266}$$

where $\rho = E[L]/E[S]$. Using (266), (263) is reduced to:

$$\phi \leq 1 - (1-f)^\rho \leq \rho f, \tag{267}$$

where $f < 1$ is given in (261), 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 (261). This immediately leads $\phi(j)$ in (262) into:

$$\frac{\phi(j)}{jf} = \frac{1-(1-f)^j}{jf} \to 1, \quad \rho \to \infty. \tag{268}$$

Invoking (268), isolation probability ϕ can be transformed into the following for $\rho \to \infty$:

$$\frac{\phi}{f\rho} = \frac{\sum_{j=1}^\infty \phi(j) P(D=j)}{f\rho} \to \frac{fE[D]}{f\rho}, \tag{269}$$

which directly leads to (259) recalling (266). □

The result in (259) indicates that for $\rho \to \infty$, probability ϕ for any user v

Table V. Comparison of the asymptotic model (258) to the exact model (255) 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%

to become isolated within its lifetime L can be approximated as the summation of probabilities that v is isolated in each individual 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 V illustrates the relative distance between the upper bound in (258) and the exact result (255) for $E[L] = 0.5$ hours and $r = 8$. It is clear from the table that as ρ increases, the two models converge and that 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.

7.3.4 Graph Disconnection

Notice that Bernoulli variable X_i in (240) 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:

$$\phi = P(X_i = 1) = 1 - P(X_i = 0), \tag{270}$$

where ϕ is given by (255) or approximated by the upper bound in (258). 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:

$$X_N = \sum_{i=1}^{N} X_i, \tag{271}$$

is the number of isolations among N join events. In the following, we use the Chen-Stein method [5] 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 [45] 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 22. *Given that $N\phi r \to 0$ as $N \to \infty$, the probability that Chord survives N user joins without disconnection approaches:*

$$\lim_{N\to\infty} \frac{P(X_N = 0)}{(1-\phi)^N} = 1, \tag{272}$$

where X_N is defined in (271) *and ϕ is given in* (255).

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 [5]:

$$|P(X_N = 0) - P(V_N = 0)| \leq \alpha(b_1 + b_2 + b_3), \tag{273}$$

where V_N is a Poisson random variable with mean $E[V_N] = E[X_N] = N\phi$, $\alpha =$

$\min(1, 1/E[X_N])$, and constants b_1, b_2 and b_3 are defined in [5]. 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:

$$B_i = \{i - r + 1, \ldots, i, \ldots, i + r - 1\} \tag{274}$$

with $i \in B_i$ and size $|B_i| = 2r - 1$. It follows that $b_3 = 0$ since Bernoulli variable X_i is independent of X_j for $j \notin B_i$. To calculate b_1, note that:

$$\begin{aligned} b_1 &= \sum_{i=1}^{N} \sum_{j \in B_i} P(X_i = 1) P(X_j = 1) = \sum_{i=1}^{N} \sum_{j \in B_i} \phi^2 \\ &= N(2r - 1)\phi^2. \end{aligned} \tag{275}$$

Likewise, we obtain:

$$\begin{aligned} b_2 &= \sum_{i=1}^{N} \sum_{j \neq i, j \in B_i} P(X_i = X_j = 1) \\ &= \sum_{i=1}^{N} \phi \sum_{j \neq i, j \in B_i} P(X_j = 1 | X_i = 1) \\ &\leq N\phi(2r - 2). \end{aligned} \tag{276}$$

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 (273) that X approaches a Poisson random variable with mean $E[X_N]$. This directly leads to:

$$\lim_{N \to \infty} \frac{P(X_N = 0)}{e^{-E[X_N]}} = \lim_{N \to \infty} \frac{P(X_N = 0)}{e^{-N\phi}} = 1. \tag{277}$$

Recalling that $\phi \to 0$ as $N \to \infty$ given the assumption of this theorem and using

Taylor expansion $e^{-\phi} = 1 - \phi$ for $\phi \to 0$, (277) yields:

$$\lim_{N\to\infty} \frac{P(X_N = 0)}{(1-\phi)^N} = 1, \tag{278}$$

which establishes the desired result. □

Theorem 22 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 non-isolated. Note that (272) 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.

Armed with (272), we propose the following approximation to $P(X_N = 0)$ for finite N:

$$P(X_N = 0) \approx (1-\phi)^N, \tag{279}$$

where the exact model of ϕ is given by (255) and its asymptotic approximation is shown in (258).

Comparison of simulation results of $P(X_N = 0)$ to (279) is presented in Table VI where model ϕ is computed based on (255). Notice from the first three columns in this table that simulation results are very close to (279) 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 (279), assume that Chord has a mean size $5,000$ users, $r = \lceil \log_2 5000 \rceil = 13$ successors, $E[L] = 0.5$ hours and $E[S] = 21$ seconds. We then obtain from (279) that the probability that Chord

Table VI. Comparison of model (279) 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.	(279)	ρ	$E[S]$ s	Simul.	(279)
$1,000$	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

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.

7.4. 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 [39], [42] to obtain Markovian models, Chord by default uses uniform intervals [80]. 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.

7.4.1 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:

$$\frac{\phi_u}{\rho f_u} \to 1, \tag{280}$$

as $E[S] \to 0$, which is similar to the result shown in (259). 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 (258). Deriving f_u, we obtain the next theorem.

Theorem 23. *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:*

$$\lim_{E[S]\to 0} \frac{\phi_u}{\phi} = \frac{2^r}{(r+1)!}. \tag{281}$$

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 (257) 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:

$$\lim_{E[S]\to 0} f_u = \frac{\rho(1-e^{-2/\rho})^{r+1}}{2(r+1)} = \frac{2^r}{(r+1)\rho^r}, \tag{282}$$

where the last step is obtained using Taylor expansion.

Next, recall from (269) 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 $D(t) \to \infty$), it is clear from the strong law of large numbers that:

$$D(t)E[S] \to t. \tag{283}$$

Invoking (283) and integrating $D(t)$ using PDF $f_L(t)$ of user lifetimes L leads to:

$$E[S]E[D] = \int_0^\infty E[S]D(t)f_L(t)dt \to E[L], \tag{284}$$

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

$$\lim_{E[S]\to 0} \frac{E[D]}{\rho} = 1 \tag{285}$$

for any distribution of S. Combining (269) and (285), we immediately obtain isolation probability ϕ_u for uniform S:

$$\frac{\phi_u}{\rho f_u} \to 1, \quad E[S] \to 0. \tag{286}$$

It is then ready to see that the ratio of ϕ_u to ϕ shown in (259) for exponential S converges to:

$$\frac{\phi_u}{\phi} \to \frac{f_u}{f}, \quad \rho \to \infty, \tag{287}$$

Table VII. Convergence of simulation results to model $\phi_u/\phi = .0127$ from (281) 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

where f is given in (258) 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 (258) can be reduced to:

$$\lim_{\rho\to\infty} f = r!\frac{e^r}{\rho^r}\left(1 - \frac{r}{\rho + r}\right)^{\rho+r+1/2} = \frac{r!}{\rho^r}, \tag{288}$$

where the last step is obtained based on Taylor expansion for for fixed r. Finally, substituting (282) and (288) into (287) directly leads to (281). □

Simulation results of ϕ_u for uniform S are shown in Table VII. Notice from this table that the ratio ϕ_u/ϕ indeed approaches that given by our model (281) as $E[S]$ becomes small. Since $\phi_u \leq \phi$ for all r, the above result demonstrates that using uniform S is a better strategy than using exponential S and that 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 $r = 12$.

7.4.2 Constant Stabilization Delays

Next, following the derivations of ϕ_u/ϕ in Theorem 23, we easily obtain isolation probability ϕ_c for constant S.

Theorem 24. *For fixed r and $E[L]$, and constant S, the ratio of isolation probability ϕ_c to ϕ approaches:*

$$\lim_{E[S]\to 0} \frac{\phi_c}{\phi} = \frac{1}{r!}. \tag{289}$$

Proof. Following the derivations in the proof for Theorem 23, we readily obtain:

$$f_c = \left(1 - e^{-1/\rho}\right)^r \to \rho^{-r}, \quad \rho \to \infty, \tag{290}$$

and

$$\frac{\phi_c}{\phi} \to \frac{f_c}{f}, \quad \rho \to \infty, \tag{291}$$

where f for exponential S is given in (288) and $\rho \to \infty$ is satisfied under given assumptions. Substituting (288) and (290) into (291) immediately leads to (289). □

Table VIII presents simulation results on ϕ_c when stabilization intervals are constant. Notice that ratio ϕ_c/ϕ obtained from simulations is very close to that predicted by model (289) even for $\rho = 60$ and that it converges to (289) as ρ increases further. Model (289) 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 (281) and (289) that $\phi_c \leq \phi_u$ and the ratio ϕ_c/ϕ_u approaches $(r+1)/2^r \leq 1$ as $E[S] \to 0$. This ratio is 0.035 for $r = 8$ and 0.003 for $r = 12$.

Table VIII. Convergence of simulation results to model $\phi_c/\phi = .0014$ from (289) 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

7.4.3 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:

$$\lim_{E[S]\to 0} \frac{\phi_c}{\phi_o} = \frac{P(\max\{L_1, \ldots, L_r\} < E[S])}{P(\max\{L_1, \ldots, L_r\} < S)}, \tag{292}$$

where $L_i \sim \exp(\mu)$ is the residual lifetime of the i-th 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 25. *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:

$$P(\max\{L_1, \ldots, L_r\} < S) = \int_0^\infty G(x) f_S(x) ds, \tag{293}$$

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

$$G''(x) = r\mu^2 e^{-\mu x}(1 - e^{-\mu x})^{r-2}(re^{-\mu x} - 1), \tag{294}$$

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

$$\begin{cases} G''(x) > 0 & x < E[L] \ln r \\ G''(x) \leq 0 & \text{otherwise} \end{cases}, \tag{295}$$

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

For $E[S] \to 0$, notice that $S \leq E[L] \ln r$ holds with probability approaching 1. This immediately transforms (293) into:

$$P(\max\{L_1, \ldots, L_r\} < S) = \int_0^{E[L] \ln r} G(x) f_S(x) ds, \tag{296}$$

showing that the convex part of $G(x)$ determines the above metric. Then, for $E[S] \to 0$ we obtain from Jensen's inequality [38] 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:

$$\lim_{E[S] \to 0} \frac{\phi_c}{\phi_o} = \frac{P(\max\{L_1, \ldots, L_r\} < E[S])}{P(\max\{L_1, \ldots, L_r\} < S)} \leq 1, \tag{297}$$

for any random S, which completes the proof. □

Theorem 25 shows that using constant S is not only a simple but *optimal* method to stabilize successors in Chord.

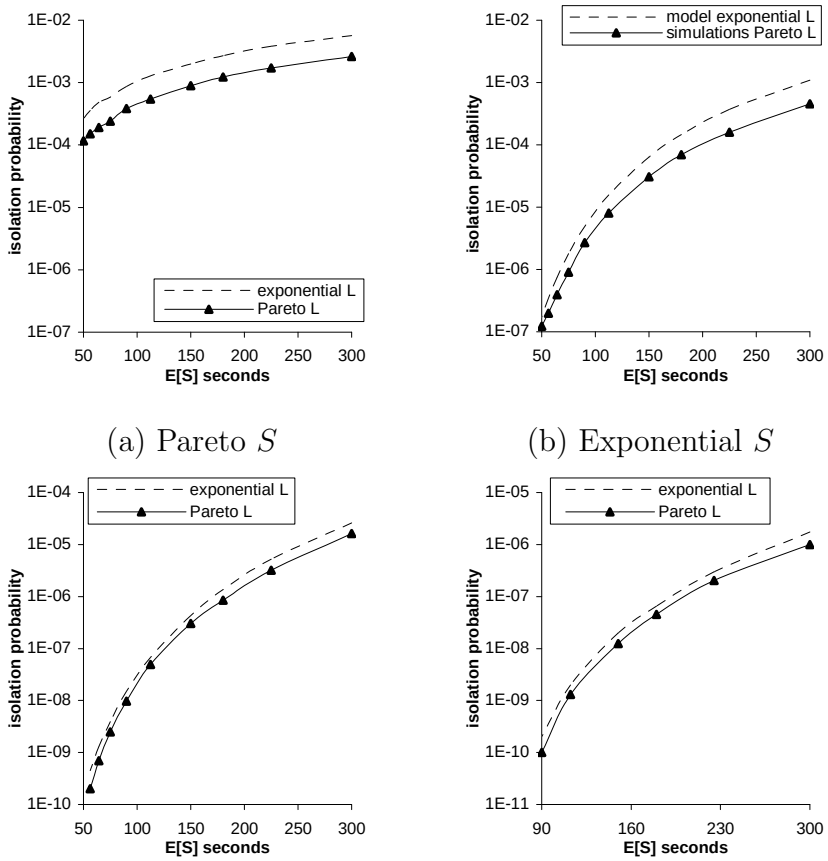
7.5. Heavy-tailed Lifetimes

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 [12], [89] 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 chapter, 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. 37. 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 that the difference becomes smaller as $E[S]$ decreases. In fact, the model is a very close match to the Pareto case in Fig. 37(c)-(d). 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.

7.6. Summary

This chapter 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



(a) Pareto S (b) Exponential S

(c) Uniform S (d) Constant S

Fig. 37. 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.

static node failure, Chord exhibited the same resilience through the successor list as that many other DHTs and unstructured P2P networks [45] through their randomized neighbor tables. We also demonstrated that when Chord experienced continuous node joins/departures, stabilization with constant intervals was optimal and kept Chord connected with the highest probability.

CHAPTER VIII

CONCLUSION AND FUTURE WORK

8.1. Conclusion

This dissertation started with proposing a novel model for user churn in P2P systems and later utilized it to understand P2P resilience under a variety of conditions on user lifetimes and graph construction. Our work can be broadly partitioned into the following five topics.

Heterogeneous Churn Model [93]. Previous analytical work has universally assumed exponential user lifetimes and homogenous users. However, measurement studies have recently revealed that user lifetimes in real P2P networks were heavy-tailed and users differed in terms of resources they contributed to the network. Our work proposed a much more generic churn model that allowed non-exponential lifetimes and captured the heterogeneous behavior of peers, including their difference in availability (i.e., the percentage of time a user is logged in), online habits, and diversity of offline delays. In this model, each user was viewed as an alternating renewal process that was ON when the user was logged in and OFF otherwise. Despite the complexity of user arrivals in this model, we showed that the aggregate lifetime distribution of joining peers was sufficient to completely characterize the effect of churn on heterogeneous P2P networks, but only when system size was asymptotically large.

Node Out-degree and Age-Based Neighbor Selection [96]. Users in *unstructured* P2P systems rely solely on their routing tables to reduce lookup latency, avoid isolation of individual nodes, and prevent graph partitioning. Prior work including our early results [43] focused on neighbor dynamics under uniform selection in net-

works with exponential user lifetimes. Our work in this part of the dissertation built a *non-exponential* model that offered exact computation of isolation probabilities for any monotone lifetime distribution, including heavy-tailed cases. The versatility of this model was illustrated by analyzing the node out-degree process under various neighbor-selection strategies in unstructured P2P networks. Leveraging the decreasing failure rate property of heavy-tailed lifetimes (i.e., larger node age means smaller failure probability) observed in real P2P networks, we proposed a novel age-proportional distributed algorithm for creating links that converged isolation probability to zero as system size became infinite.

Node In-Degree [93]. The above approach focused on only out-degree neighbors and did not consider the impact of in-degree neighbors on resilience. We formally proved that under heterogeneous user churn and uniform neighbor selection, the edge-arrival process to each user approached Poisson as system size became sufficiently large. This led us to simple analytical treatment of the edge-arrival process and offered closed-form results on the transient distribution of in-degree as a function of the aggregate user lifetime distribution and clearly illustrated the contribution of in-degree to resilience.

Link Lifetimes in DHTs [94]. Several models of user churn, resilience, and link lifetime have recently appeared in the literature [42], [45], [93]; however, these results do not directly apply to classical Distributed Hash Tables (DHTs) in which neighbor replacement occurs not only when current users die, but also when new users arrive into the system, and where replacement choices are often restricted to the successor of the failed zone in the DHT space. Using a semi-Markov chain, we showed that the zone size (i.e., fraction of the DHT key space) of neighbors plays a crucial role in link lifetimes and proposed a min-zone algorithm to significantly improve the resilience of DHTs to node isolation.

Successor Lists in DHTs [95]. Previous analytical work [42], [45] 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 part, 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.

8.2. Future Work

Future work includes derivation of residual lifetime distributions in finite systems, development of more sophisticated algorithms for increased DHT resilience, and analysis of neighbor selection techniques in asymptotically small networks where limiting results similar to Theorem 19 do not hold.

The other direction involves modeling *non-stationary* user churn in P2P networks. Despite the elegance and pervasive use of *stationary* models in prior work including ours, measurement studies have revealed that user churn in P2P systems was non-stationary. In fact, non-stationary churn models are applicable to many user-driven systems, where time-varying arrival/departure processes reflect the rhythm of human activity. Future work includes offering generic non-Poisson models that can be applied to a broader class of problems, analysis of the performance of networked systems under churn, and verification of theoretical results in real networks.

REFERENCES

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

[2] R. Albert and A. Barabási, "Topology of Evolving Networks: Local Events and Universality," *Phys. Rev. Lett.*, vol. 85, no. 24, pp. 5234–5237, Dec. 2000.

[3] R. Albert, H. Jeong, and A. L. Barabási, "Error and Attack Tolerance of Complex Networks," *Nature*, vol. 406, pp. 378–382, 2000.

[4] D. Aldous and M. Brown, "Inequalities for Rare Events in Time-Reversible Markov Chains I," in *Stochastic Inequalities*, M. Shaked and Y. L. Tong, Eds. Hayward, CA: Institute of Mathematical Statistics, 1992, vol. 22, pp. 1–16.

[5] 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.

[6] J. Aspnes, Z. Diamadi, and G. Shah, "Fault-Tolerant Routing in Peer-to-Peer Systems," in *Proc. ACM PODC*, Jul. 2002, pp. 223–232.

[7] N. Balakrishnan and M. V. Koutras, *Runs and Scans with Applications.* New York, NY: John Wiley & Sons, 2002.

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

[9] BitTorrent. (2007, Apr.). [Online]. Available: http://www.bittorrent.com.

[10] A. A. Borovkov, *Probability Theory.* New York, NY: Gordon and Breach, 1998.

[11] J. T. Bradley, N. J. Dingle, P. G. Harrison, and W. J. Knottenbelt, "Distributed Computation of Passage Time Quantiles and Transient State Distributions in

Large Semi-Markov Models," in *Proc. IPDPS*, Apr. 2003.

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

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

[14] E. Çinlar, *Introduction to Stochastic Processes*. Englewood Cliffs, NJ: Prentice Hall, 1997.

[15] Y. Chawathe, S. Ratnasamy, L. Breslau, N. Lanham, and S. Shenker, "Making Gnutella-like P2P Systems Scalable," in *Proc. ACM SIGCOMM*, Aug. 2003, pp. 407–418.

[16] B.-G. Chun, B. Zhao, and J. Kubiatowicz, "Impact of Neighbor Selection on Performance and Resilience of Structured P2P Networks," in *Proc. IPTPS*, Feb. 2005, pp. 264–274.

[17] L. Devroye, "Law of the Iterated Logarithm for Order Statistics of Uniform Spacings," *Annals of Probability*, vol. 9, pp. 860–867, 1981.

[18] E. B. Dynkin, "Some Limit Theorems for Sums of Independent Random Variables with Infinite Mathematical Expectations," *Selected Transl. in Math. Statist. and Probab.*, vol. 1, pp. 171–189, 1961.

[19] H. Exton, *Handbook of Hypergeometric Integrals: Theory, Applications, Tables, Computer Programs*. Chichester, U.K.: Ellis Horwood, 1978.

[20] A. Feldmann and W. Whitt, "Fitting Mixtures of Exponentials to Long-tailed Distributions to Analyze Network Performance Models," *Performance Evaluation*, vol. 31, no. 3-4, pp. 245–279, Jan. 1998.

[21] W. Feller, *An Introduction to Probability Theory and Its Applications, Volume 2.* New York, NY: John Wiley & Sons, 1966.

[22] H. Frank, "Maximally Reliable Node Weighted Graphs," in *Proc. 3rd Ann. Conf. Information Sciences and Systems*, Mar. 1969, pp. 1–6.

[23] C. Gkantsidis, M. Mihail, and A. Saberi, "Random Walks in Peer-to-Peer Networks," in *Proc. IEEE INFOCOM*, Mar. 2004, pp. 120–130.

[24] C. Gkantsidis, M. Mihail, and A. Saberi, "Hybrid Search Schemes for Unstructured Peer-to-Peer Networks," in *Proc. IEEE INFOCOM*, Mar. 2005, pp. 1526–1537.

[25] Gnutella. (2007, Mar.). [Online]. Available: http://www.gnutella.com/.

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

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

[28] S. Guha, N. Daswani, and R. Jain, "An Experimental Study of the Skype Peer-to-Peer VoIP System," in *Proc. IPTPS*, 2006.

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

[30] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan, "Measurement, Modeling, and Analysis of a Peer-to-Peer File-Sharing Workload," in *Proc. ACM SOSP*, Oct. 2003, pp. 314–329.

[31] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang, "Measurements, Analysis, and Modeling of Bittorrent-Like Systems," in *Proc. ACM IMC*, 2005.

[32] T. Hettmansperger and M. Keenan, "Tailweight, Statistical Inference, and Families of Distributions– A Brief Survey," in *Statist. Distributions in Scientific Work*, G. P. Patil et al., Ed. Boston, MA: Kluwer, 1980, vol. 1, pp. 161–172.

[33] S. Ioannidis and P. Marbach, "On the Design of Hybrid Peer-to-Peer Systems," in *Proc. ACM SIGMETRICS*, Jun. 2008, pp. 157–168.

[34] M. F. Kaashoek and D. Karger, "Koorde: A Simple Degree-Optimal Distributed Hash Table," in *Proc. IPTPS*, Feb. 2003, pp. 98–107.

[35] KaZaA. (2008, Jan.). [Online]. Available: http://www.kazaa.com/.

[36] A. K. Kelmans, "Connectivity of Probabilistic Networks," *Auto. Remote Contr.*, vol. 29, pp. 444–460, 1967.

[37] M. Kijima, *Markov Processes for Stochastic Modeling*. London, U.K.: Chapman & Hall, 1997.

[38] S. G. Krantz, *Handbook of Complex Variables*. Boston, MA: Birkhäuser, 1999.

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

[40] S. S. Lam and H. Liu, "Failure Recovery for Structured P2P Networks: Protocol Design and Performance Evaluation," in *Proc. ACM SIGMETRICS*, Jun. 2004, pp. 199–210.

[41] J. Ledlie, J. Shneidman, M. Amis, and M. Seltzer, "Reliability- and Capacity-Based Selection in Distributed Hash Tables," Harvard University Computer Science, Tech. Rep., Sep. 2003.

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

[43] D. Leonard, Z. Yao, V. Rai, and D. Loguinov, "On Lifetime-Based Node Failure and Stochastic Resilience of Decentralized Peer-to-Peer Networks," *IEEE/ACM Trans. Networking*, vol. 15, no. 3, pp. 644–656, Jun. 2007.

[44] D. Leonard, Z. Yao, X. Wang, and D. Loguinov, "On Static and Dynamic Partitioning Behavior of Large-Scale P2P Networks," *IEEE/ACM Trans. Networking*, vol. 16, no. 6, pp. 1475–1488, Dec. 2008.

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

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

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

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

[49] J. Liang, R. Kumar, and K. W. Ross, "The KaZaA Overlay: A Measurement Study," *Computer Networks*, 2005.

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

[51] D. Loguinov, J. Casas, and X. Wang, "Graph-Theoretic Analysis of Structured Peer-to-Peer Systems: Routing Distances and Fault Resilience," *IEEE/ACM*

Trans. Networking, vol. 13, no. 5, pp. 1107–1120, Oct. 2005.

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

[53] L. Lovász, "Random Walks on Graphs: A Survey," in *Combinatorics, Paul Erdös is Eighty*, D. Miklós et al., Ed. Budapest, Hungary: János Bolyai Mathematical Society, 1996, vol. 2, pp. 353–398.

[54] E. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim, "A Survey and Comparison of Peer-to-Peer Overlay Network Schemes," *IEEE Communications Surveys & Tutorials*, vol. 7, no. 2, pp. 72–93, 2005.

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

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

[57] L. Massoulié, A.-M. Kermarrec, and A. Ganesh, "Network Awareness and Failure Resilience in Self-Organising Overlay Networks," in *Proc. IEEE SRDS*, Oct. 2003, pp. 47–55.

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

[59] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA: Society for Industrial and Applied Math, 2000.

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

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

[62] V. V. Petrov, *Sums of Independent Random Variables*. New York, NY: Springer-Verlag, 1975.

[63] C. G. Plaxton, R. Rajaraman, and A. W. Richa, "Accessing Nearby Copies of Replicated Objects in a Distributed Environment," in *Proc. ACM SPAA*, 1997, pp. 311–320.

[64] L. Plissonneau, J.-L. Costeux, and P. Brown, "Analysis of Peer-to-Peer Traffic on ADSL," in *Proc. PAM*, Mar. 2005, pp. 69–82.

[65] J. Pouwelse, P. Garbacki, D. Epema, and H. Sips, "The Bittorrent P2P File-Sharing System: Measurements and Analysis," in *Proc. IPTPS*, 2005.

[66] D. Qiu and R. Srikant, "Modeling and Performance Analysis of BitTorrent-Like Peer-to-Peer Networks," in *Proc. ACM SIGCOMM*, Aug. 2004, pp. 367–378.

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

[68] S. Ratnasamy, M. Handley, R. Karp, and S. Shenker, "Topologically-Aware Overlay Construction and Server Selection," in *Proc. IEEE INFOCOM*, Jun. 2002, pp. 1190–1199.

[69] S. I. Resnick, *Extreme Values, Regular Variation, and Point Processes*. New York, NY: Springer-Verlag, 1987.

[70] S. I. Resnick, *Adventures in Stochastic Processes*. Boston, MA: Birkhäuser, 2002.

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

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

[73] D. Rubenstein and S. Sahu, "Can Unstructured P2P Protocols Survive Flash Crowds," *IEEE/ACM Trans. Netw.*, vol. 13, no. 3, pp. 501–512, Apr. 2005.

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

[75] S. Saroiu, P. K. Gummadi, and S. D. Gribble, "Analyzing the Characteristics of Napster and Gnutella Hosts," *Multimedia Systems*, vol. 9, pp. 170–184, 2003.

[76] Skype. (2008, Nov.). [Online]. Available: http://www.skype.com.

[77] K. Sripanidkulchai, A. Ganjam, B. Maggs, and H. Zhang, "The Feasibility of Supporting Large-Scale Live Streaming Applications with Dynamic Application End-Points," in *Proc. ACM SIGCOMM*, Aug. 2004, pp. 107–120.

[78] M. Srivatsa, B. Gedik, and L. Liu, "Large Scaling Unstructured Peer-to-Peer Networks with Heterogeneity-Aware Topology and Routing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 17, no. 11, pp. 1277–1293, Nov. 2006.

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

[80] 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. Netw.*, vol. 11, no. 1, pp. 17–32, Feb. 2003.

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

[82] D. Stutzbach, R. Rejaie, and S. Sen, "Characterizing Unstructured Overlay Topologies in Modern P2P File-Sharing Systems," in *Proc. ACM IMC*, Oct. 2005, pp. 49–62.

[83] K. Sutner, A. Satyanarayana, and C. Suffel, "The Complexity of the Residual Node Connectedness Reliability Problem," *SIAM J. Comput.*, vol. 20, pp. 149–155, 1991.

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

[85] M. S. Taqqu, W. Willinger, and R. Sherman, "Proof of a Fundamental Result in Self-Similar Traffic Modeling," *ACM Comput. Commun. Rev.*, vol. 27, no. 2, pp. 5–23, Apr. 1997.

[86] D. Towsley, "The Internet is Flat: A Brief History of Networking in the Next Ten Years," in *Proc. ACM PODC*, Aug. 2008, pp. 11–12.

[87] V. Venkataraman, P. Francisy, and J. Calandrino, "Chunkyspread: Multitree Unstructured Peer-to-Peer Multicast," in *Proc. IPTPS*, Feb. 2006.

[88] V. Vishnumurthy and P. Francis, "On Heterogeneous Overlay Construction and Random Node Selection in Unstructured P2P Networks," in *Proc. IEEE INFOCOM*, Apr. 2006.

[89] X. Wang, Z. Yao, and D. Loguinov, "Residual-Based Measurement of Peer and

Link Lifetimes in Gnutella Networks," in *Proc. IEEE INFOCOM*, May 2007, pp. 391–399.

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

[91] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[92] J. Xu, A. Kumar, and X. Yu, "On the Fundamental Tradeoffs between Routing Table Size and Network Diameter in Peer-to-Peer Networks," *IEEE J. Sel. Areas Commun.*, vol. 22, pp. 151–163, 2004.

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

[94] Z. Yao and D. Loguinov, "Link Lifetimes and Randomized Neighbor Selection in DHTs," in *Proc. IEEE INFOCOM*, Apr. 2008.

[95] Z. Yao and D. Loguinov, "Understanding Disconnection and Stabilization of Chord," in *Proc. IEEE INFOCOM*, Apr. 2008.

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

[97] H. Zhang, A. Goal, and R. Govindan, "Incrementally Improving Lookup Latency in Distributed Hash Table Systems," in *Proc. ACM SIGMETRICS*, Jun. 2003, pp. 114–125.

[98] B. Y. Zhao, L. Huang, J. Stribling, S. C. Rhea, A. D. Joseph, and J. Kubia-

towicz, "Tapestry: A Resilient Global-Scale Overlay for Service Deployment," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 1, pp. 41–53, Jan. 2004.

[99] M. Zhong, K. Shen, and J. Seiferas, "Non-Uniform Random Membership Management in Peer-to-Peer Networks," in *Proc. IEEE INFOCOM*, Mar. 2005, pp. 1151–1161.

[100] D. Zhou, J. Huang, and B. Schölkopf, "Learning from Labeled and Unlabeled Data on a Directed Graph," in *Proc. ICML 2005*, Aug. 2005, pp. 1036–1043.

VITA

Zhongmei Yao received her B.S. degree (with honors) in engineering from Donghua University, Shanghai, China, in 1997 and her M.S. degree in computer science from Louisiana Tech University, Ruston, in 2004.

She joined the Internet Research Lab in the Department of Computer Science and Engineering at Texas A&M University in January 2005 and graduated with her Ph.D. in Computer Science in August 2009. Her current research interests are in computer networking, with a focus on network modeling, stochastic process theory, algorithm design, and peer-to-peer networks. She can be reached at:

Zhongmei Yao

Department of Computer Science and Engineering

Texas A&M University

College Station, TX 77843-3112

The typist for this dissertation was Zhongmei Yao.