Modeling Heterogeneous User Churn and Local Resilience of Unstructured P2P Networks

Zhongmei Yao

Joint work with Derek Leonard, Xiaoming Wang, and Dmitri Loguinov

Internet Research Lab
Department of Computer Science
Texas A&M University, College Station, TX 77843

Nov. 13, 2006

Modeling Heterogeneous User Churn and Local Resilience of Unstructured P2P Networks

User arrival and departure process

Different users exhibit different online/offline behavior

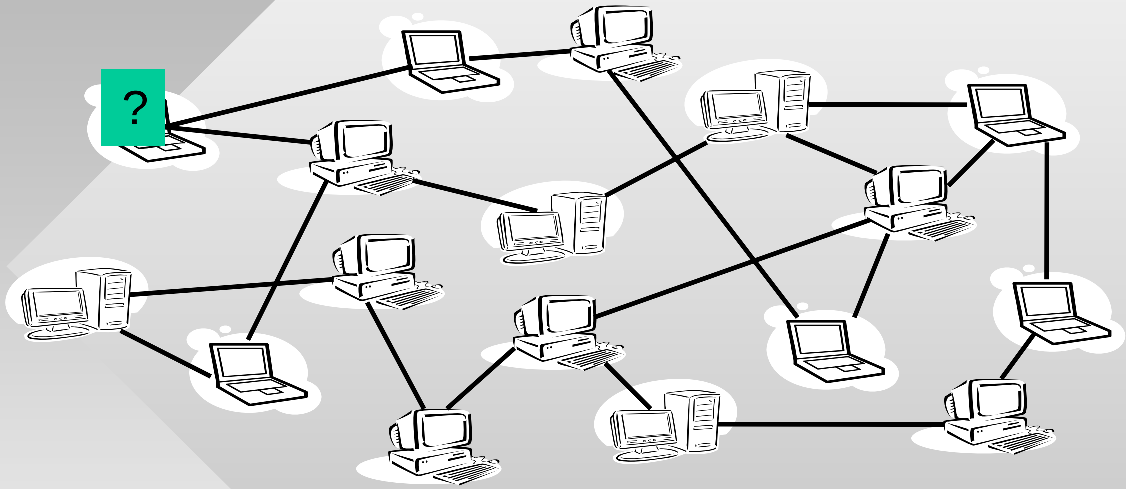
Random graphs without a-priori structure

Probability of peer isolation within his/her lifetime

Agenda

- Motivation and background
  - Terminology, assumptions, and previous work
- Heterogeneous churn model
  - Lifetime distribution of joining users
  - Residual lifetime distribution
  - Lifetime distribution of users in the system
- In-degree results (summary)
- Joint in/out-degree results (summary)
- Wrap up

P2P Networks



- Unstructured P2P networks organize peers into decentralized random graphs (Gnutella, KaZaA)
  - Search performed by routing between neighbors
- Performance depends on the state of neighboring nodes and ability of the system to stay connected during churn

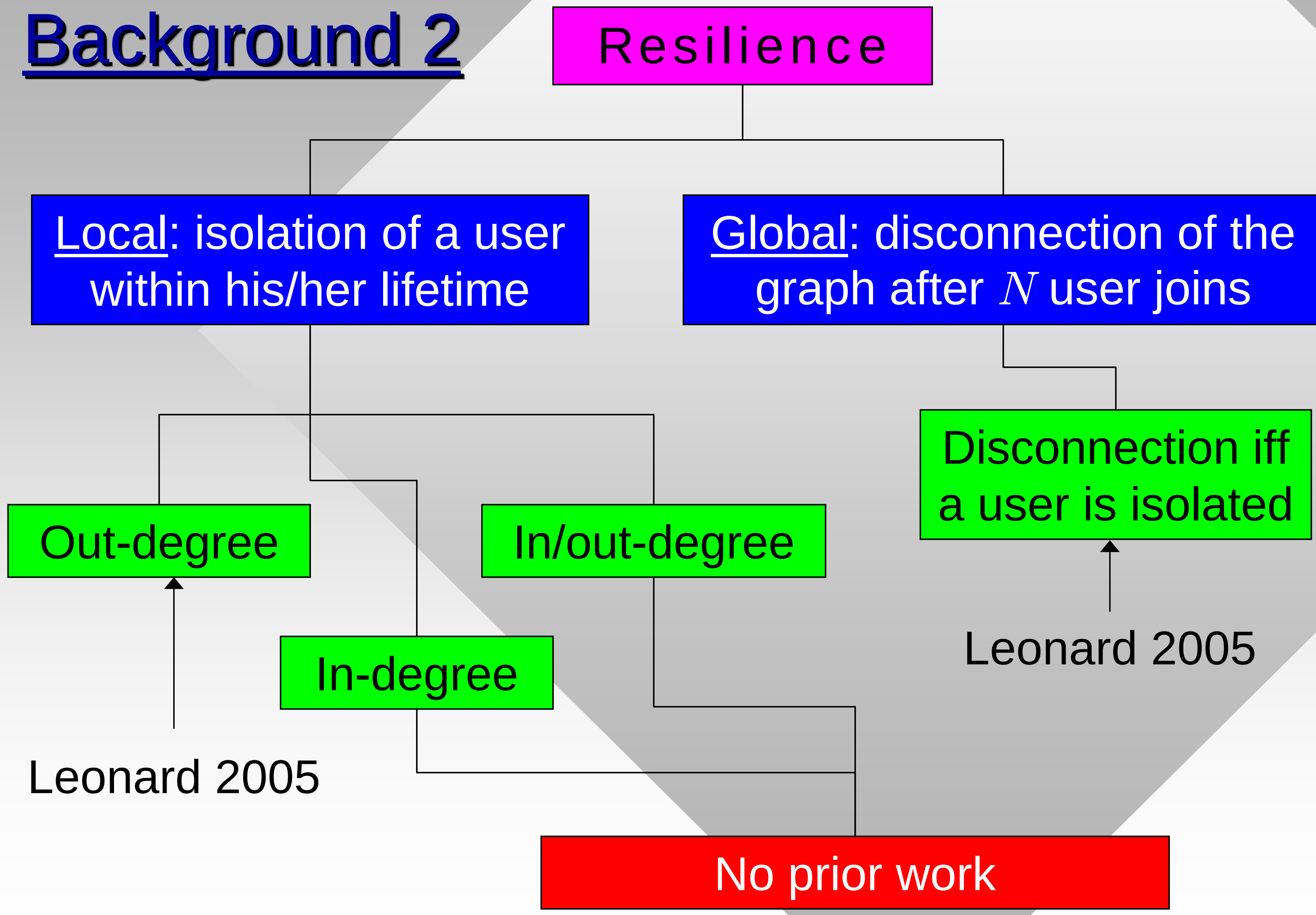
Terminology

- Churn model:
  - Arrival instances and lifetime distribution of users (no need for an explicit departure process)
- Edge creation:
  - Joining users select k random peers from the system
  - These are called out-degree neighbors
  - Users attaching to a node are its in-degree neighbors
- Replacement of neighbors:
  - Detection of failed neighbors and replacement with alive peers within S time units (can be fixed or random)
- Only out-degree neighbors are replaced to avoid unlimited degree expansion

Background

- Churn
  - Homogeneous: all users have the same distribution of lifetime
    - Exponential lifetimes
      - Independent Poisson arrivals ← Pandurangan 2001, Liben-Nowell 2002, Krishnamurthy 2005
    - Arbitrary lifetimes
      - Arrivals not modeled ← Leonard 2005
  - Heterogeneous: each user has a different online and offline distribution
    - No prior work

Background 2



Motivation

- User heterogeneity is a fundamental property of human-based networks
  - Some users consistently spend minutes logged in, others hours or even days
  - Each user’s lifetime is drawn from a user-specific distribution that describes his/her online behavior
- Churn in such networks is characterized by the distribution of both online and offline durations
  - Online/offline distributions define peer availability
- Finally, understanding of isolation and effects of churn requires in-degree characterization

Our Contributions

- Main results
  - User churn model
    - Lifetime of joining users
    - Residual lifetimes
    - Lifetime of alive users
  - In-degree model
    - User arrival process
    - Edge arrival process
    - In-degree evolution
  - Joint in/out-degree model
    - Node isolation

Agenda

- Introduction
  - Peer-to-peer networks, previous work, our main results
- Heterogeneous churn model
  - Lifetime distribution of joining users
  - Residual lifetime distribution
  - Lifetime distribution of users in the system
- In-degree model (summary)
- Joint in/out-degree (summary)
- Wrap up

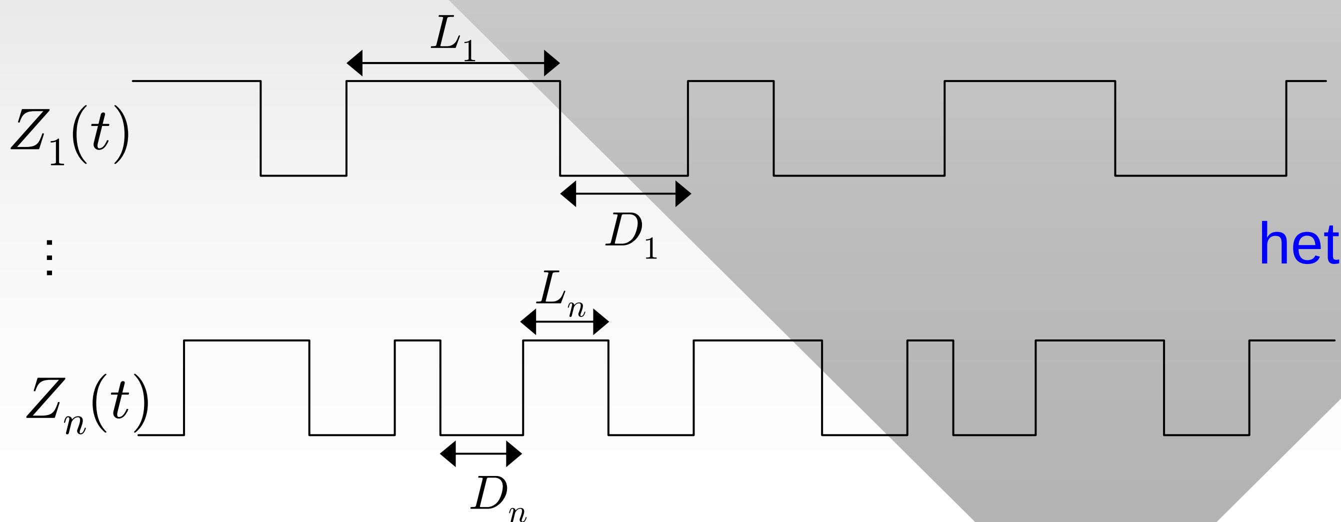
Heterogeneous User Churn

number of all possible users

- Each user's ON/OFF behavior is modeled by an alternating renewal process $\{Z_i(t)\}$

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

- ON periods $L_i \sim F_i(x)$, OFF periods $D_i \sim G_i(x)$



heterogeneous

System Population

average lifetime of user i

- User availability is defined as the long-term fraction of time a user is logged in

$$a_i = \lim_{t \to \infty} P(Z_i(t) = 1) = \frac{E[L_i]}{E[L_i] + E[D_i]}$$

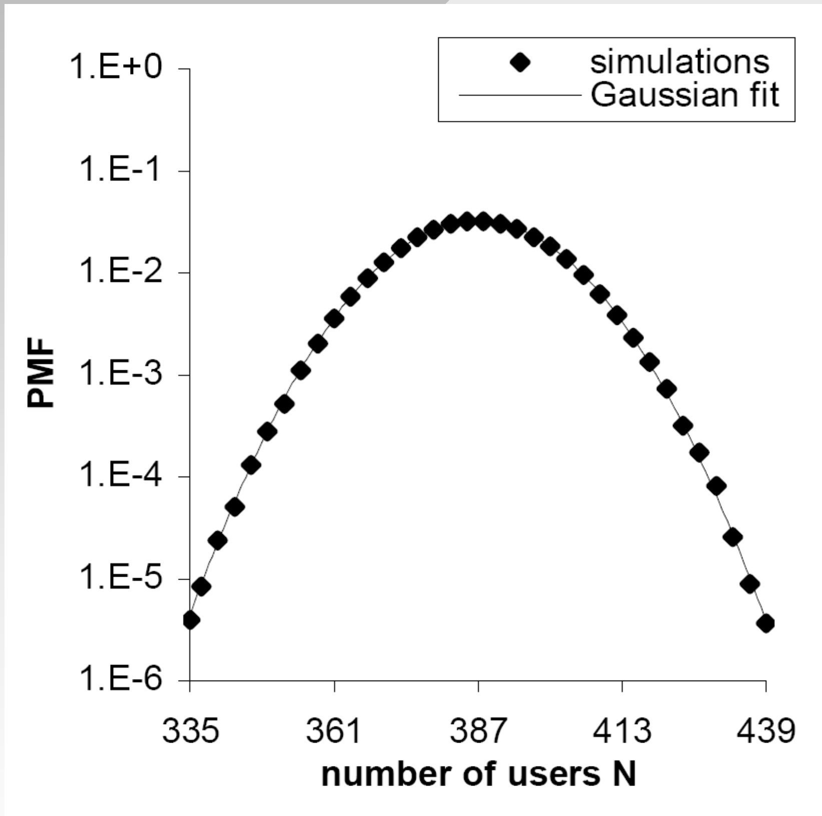
- System population at random time t is:

$$N(t) = \sum_{i=1}^{n} Z_i(t)$$

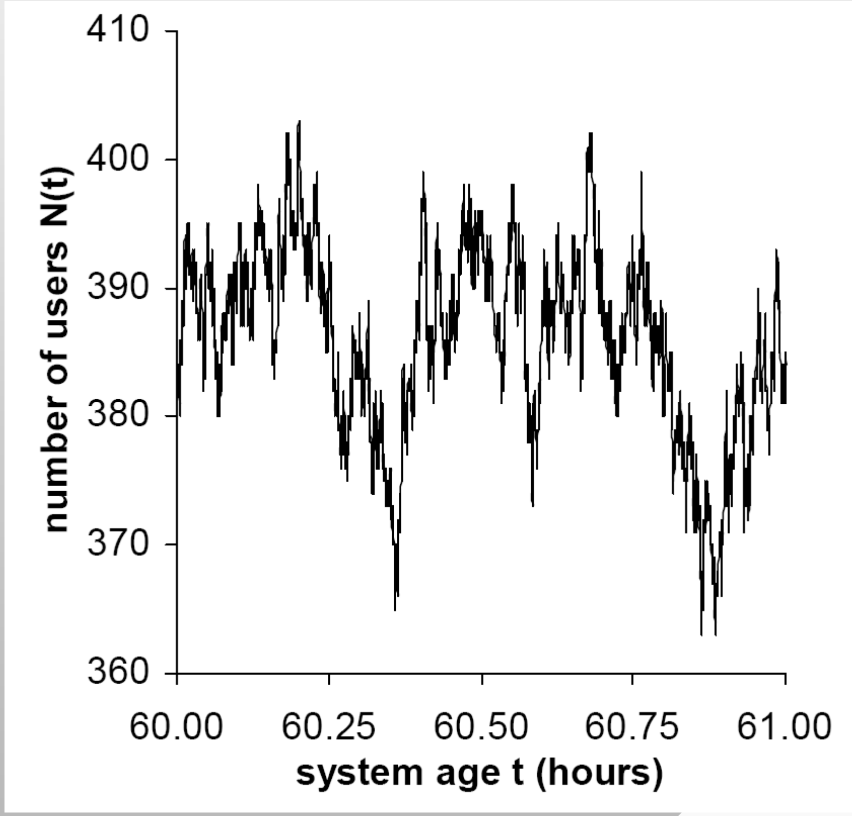
- Theorem 1: The number of users observed in the equilibrium tends to a Gaussian random variable $N(\mu, \sigma^2)$ as n approaches ∞, where:

$$\mu = \sum_{i=1}^{n} a_i, \qquad \sigma^2 = \sum_{i=1}^{n} a_i(1 - a_i)$$

System Population



(a) $N(t)$ at time t is Gaussian



(b) $\{N(t): t\geq 0\}$ is Brownian motion

Lifetime Distribution of Joining Users

- Theorem 2: The distribution of lifetime L of joining users is given by:

$$F(x) = P(L < x) = \sum_{i=1}^{n} b_i F_i(x)$$

lifetime CDF of user i

where:

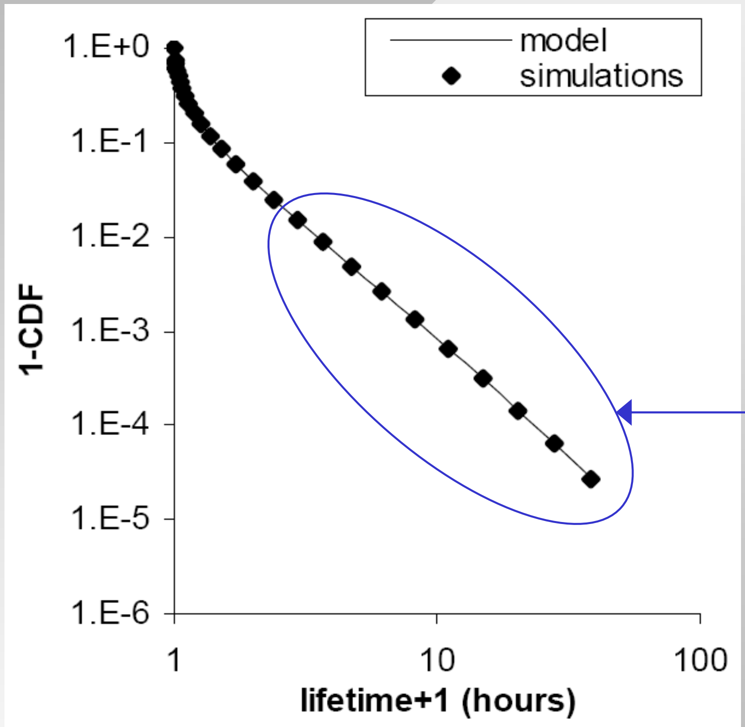
$$b_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}, \qquad \lambda_i = \frac{1}{E[L_i] + E[D_i]}$$

focus of prior measurement studies

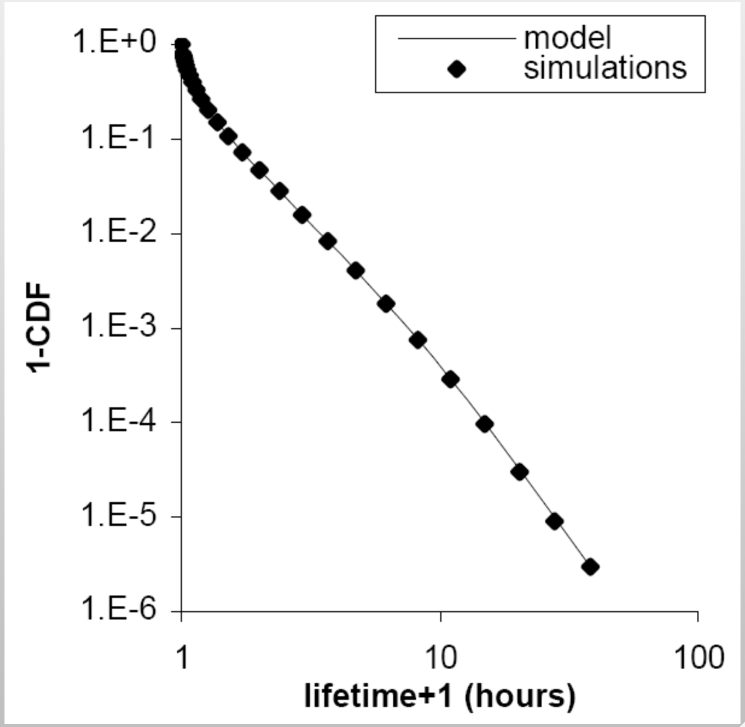
arrival rate of user i

- Weights b_i are biased toward those peers who frequently join and leave the system
  - Note that $F(x)$ is a complex mixture of individual CDFs

Lifetime Distribution of Joining Users



(a) exponential $F_i(x)$



(b) Pareto $F_i(x)$

- Aggregate lifetime distribution $F(x)$ may be heavy-tailed even if individual $F_i(x)$ are not

Lifetime Distribution of Joining Users

- For exponential $F_i(x)$, there exists a set of weights $\{b_1, \ldots, b_n\}$ such that their weighted sum converges to any monotonic distribution $W(x)$

$$\sum_{i=1}^{n} b_i F_i(x) \to W(x), \quad \text{as } n \to \infty$$

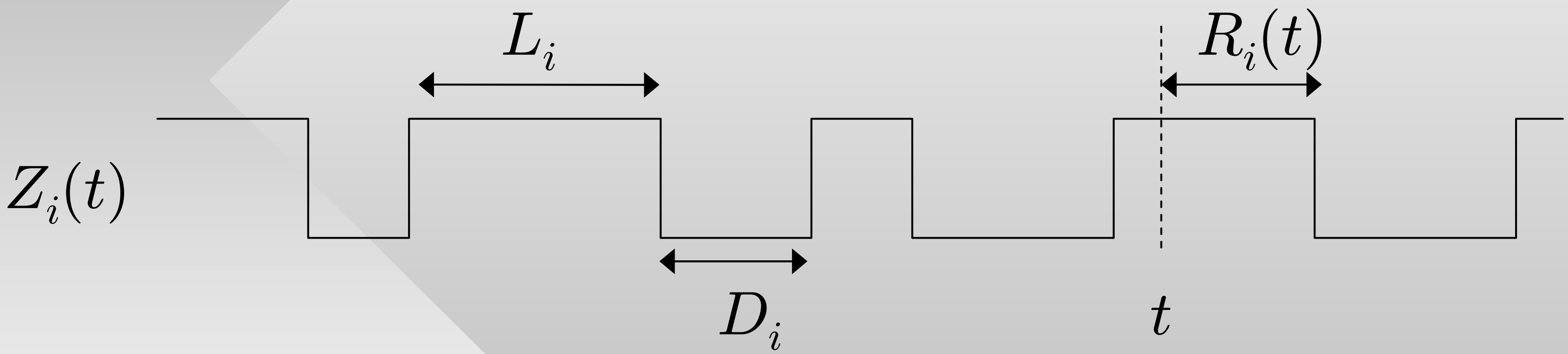
any desired distribution with a monotonic PDF

  - Depending on arrival-rate set $\{\lambda_1, \ldots, \lambda_n\}$, $W(x)$ can be Pareto, Weibull, or other distribution

- Thus, for a known aggregate distribution $F(x)$, one cannot conclude if individual user behavior bears the same nature as $F(x)$

Residual Lifetime Distribution

- Residual lifetime $R_i(t)$ of given user i is his/her remaining online duration from time t



- Let $R(t)$ be the residual lifetime of a user randomly selected by the network at time t
  - Denote its equilibrium distribution by

$$H(x) = \lim_{t \to \infty} P(R(t) < x)$$

- This metric depends on neighbor-selection strategies

Residual Lifetime Distribution

- Define the probability that user i is selected from among j alive users:

user i is alive at t

$$s_{ij} = \lim_{t \to \infty} P(i \text{ selected} | Z_i(t) = 1, N(t) = j)$$

  – Recall that individual users may have a different probability of being selected due to heterogeneity

- For uniform selection, $s_{ij} = 1/j$

degree of user i

- Using stationary random walks, $s_{ij} = d_i / \sum_{m=1}^{j} d_m$
- Under content-based selection, $s_{ij} = w_i / \sum_{m=1}^{j} w_m$

content of user i

Computer Science, Texas A&M University

Residual Lifetime Distribution

- Theorem 3: In an equilibrium system, the residual lifetime distribution of a random neighbor is given by

$$H(x) = \sum_{i=1}^{n} V_i(x) a_i \sum_{j=1}^{n} s_{ij} P(N(n-1) = j-1)$$

residual lifetime of user i condition on it being selected

availability

probability of i being selected among j alive users

- For age-independent (Leonard 2005) selection, $V_i(x)$ is the residual lifetime distribution $H_i(x)$
  - For all other cases, understanding neighbor resilience is a much more complex issue

Residual Lifetime Distribution

- Distribution of $R(t)$ involves a number of complex factors:
  - Distribution of system population $N(t)$
  - Residual lifetime distribution $V_i(x)$ of selected neighbors
  - Distribution of individual lifetimes $F_i(x)$
  - Selection strategy s_{ij}
- Analysis of residual lifetime distribution $H(x)$ is intractable unless some assumptions are made
  - From this point, we assume uniform selection that is implemented using special random walks on the graph (Zhong 2005)

Residual Lifetime Distribution

- Theorem 4: Under uniform selection, the equilibrium residual distribution $H_U(x)$ of random neighbors can be reduced to the following:

$$H_U(x) = \frac{1}{E[L]} \int_0^x (1 - F(u))du$$

where:

$$E[L] = \sum_{i=1}^{n} b_i E[L_i]$$

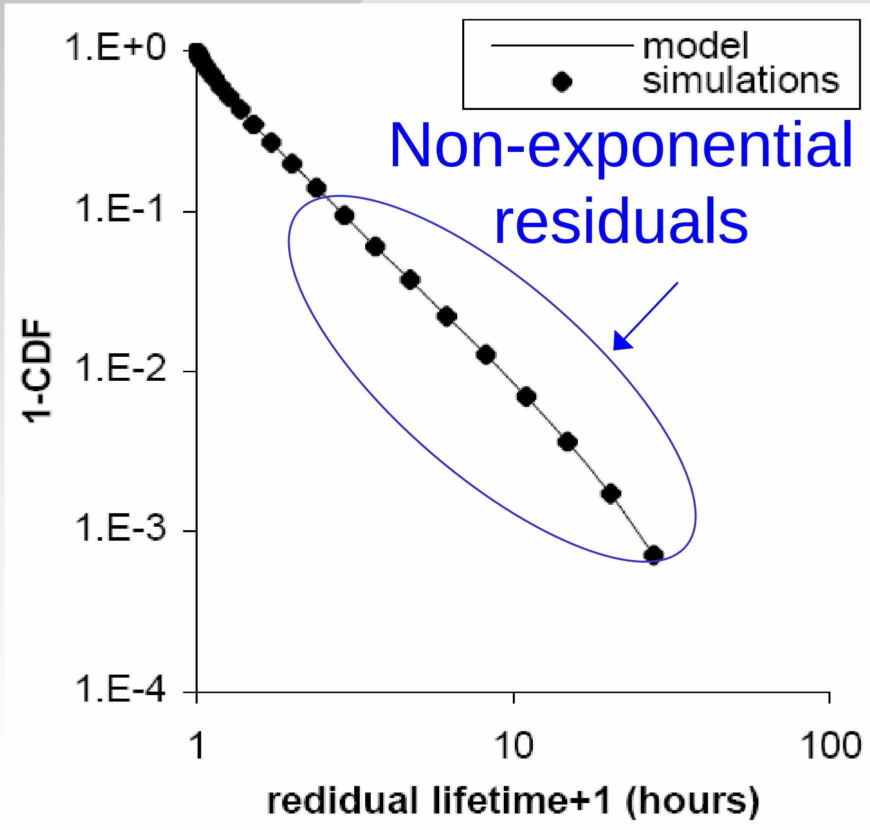
lifetime distribution of joining users

average session time of a joining user

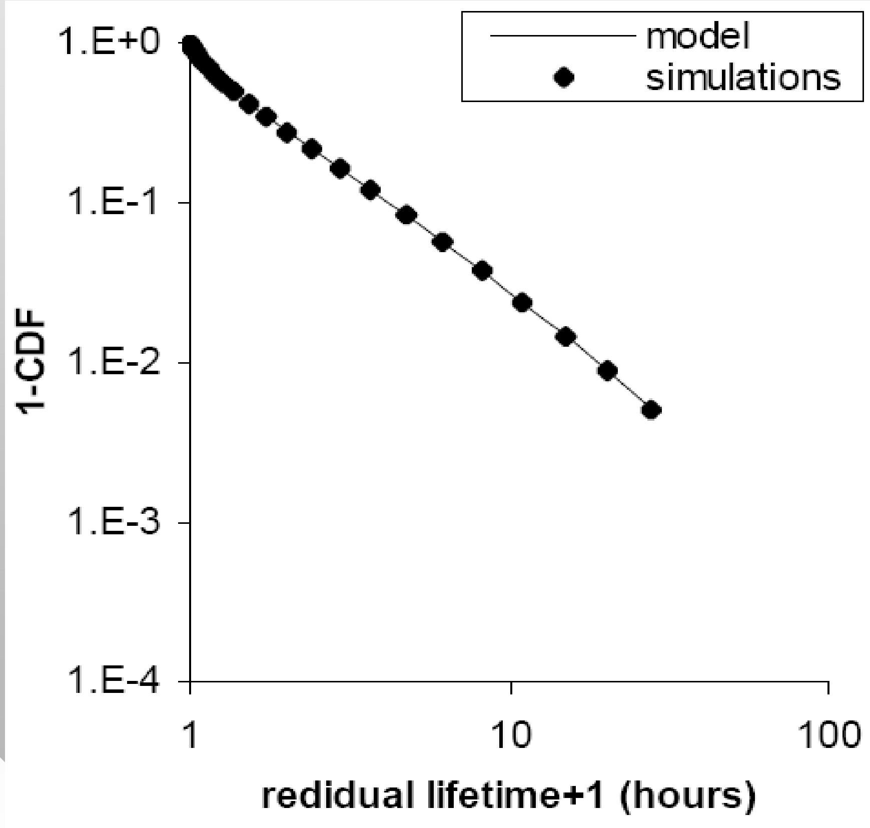
- Both $F(x)$ and $E[L]$ are easily measurable in existing systems

Residual Lifetime Distribution

- Simulation results when uniform selection is used



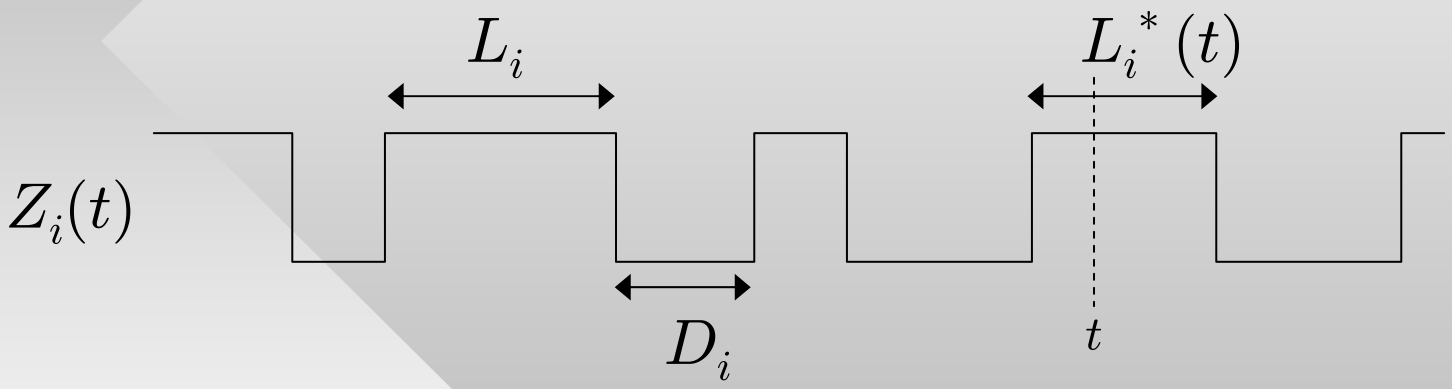
(a) exponential $F_i(x)$



(b) Pareto $F_i(x)$

Lifetime Distribution of Users in the System

- Denote by $L_i^*(t)$ the lifetime of randomly selected user i currently in the system at some time t



- Inspection paradox:
  - Lifetimes of the peers observed in the system are biased towards larger values

Lifetime Distribution of Users in the System

- Theorem 5: The joint lifetime distribution $J(x)$ of existing users in the system is:

lifetime distribution of joining users

$$J(x) = \frac{1}{E[L]}\left(xF(x) - \int_0^x F(u)du\right)$$

Furthermore, distribution $J(x)$ is the convolution of two residual lifetime distributions $H_U(x)$ and the mean lifetime of an alive user is double the mean residual lifetime of a uniformly selected peer

- Prior measurement studies have observed this difference, but it is formalized here for the first time

Discussion

- Under uniform selection, lifetimes of joining users given by CDF $F(x)$ characterizes all other related distributions and metrics
  - Instead of measuring individual user lifetimes, it is sufficient to sample lifetimes of joining peers to characterize churn
- Aggregate behavior $F(x)$ does not necessarily convey any information about individual peer lifetimes $F_i(x)$
  - Heavy-tailed $F(x)$ observed in practice does not imply individual lifetimes are heavy-tailed as well
- If selection is not uniform, our results show that the system is extremely complex and neighbor residual lifetimes are currently not tractable!

Agenda

- Motivation and background
  - Terminology, assumptions, and previous work
- Heterogeneous churn model
  - Lifetime distribution of joining users
  - Residual lifetime distribution
  - Lifetime distribution of users in the system
- In-degree results (summary)
- Joint in/out-degree results (summary)
- Wrap up

User Arrival Process

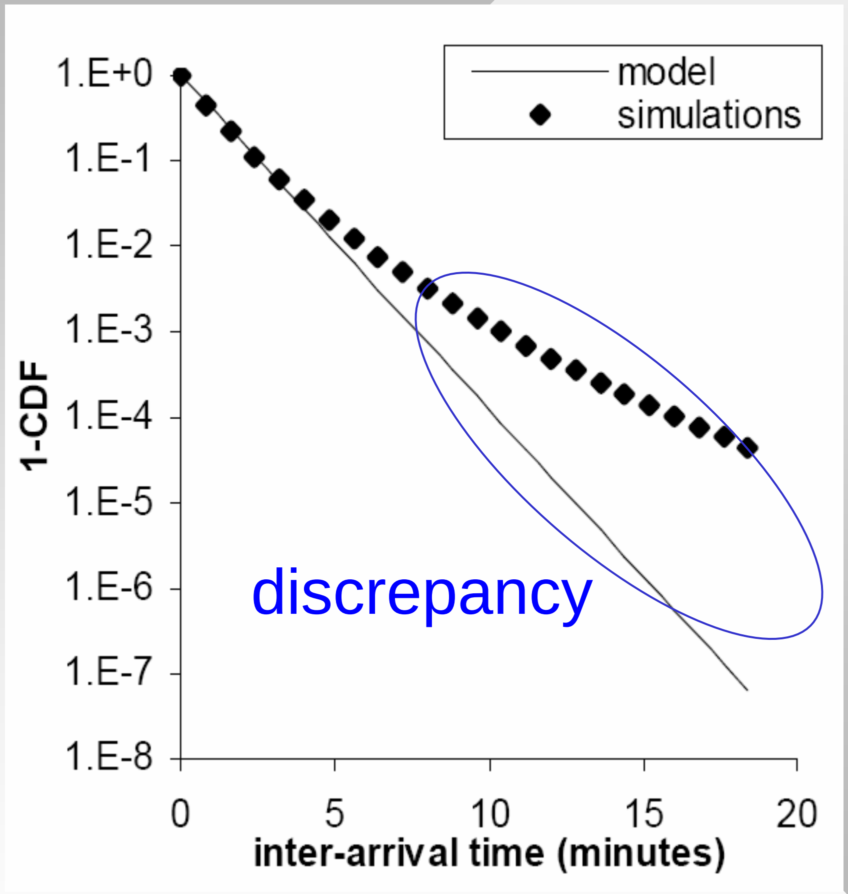
- Theorem 6: Under heterogeneous churn, user arrivals into the system converge as $n \to \infty$ to a homogeneous Poisson process with constant rate:

$$\lambda = \sum_{i=1}^{n} \lambda_i$$

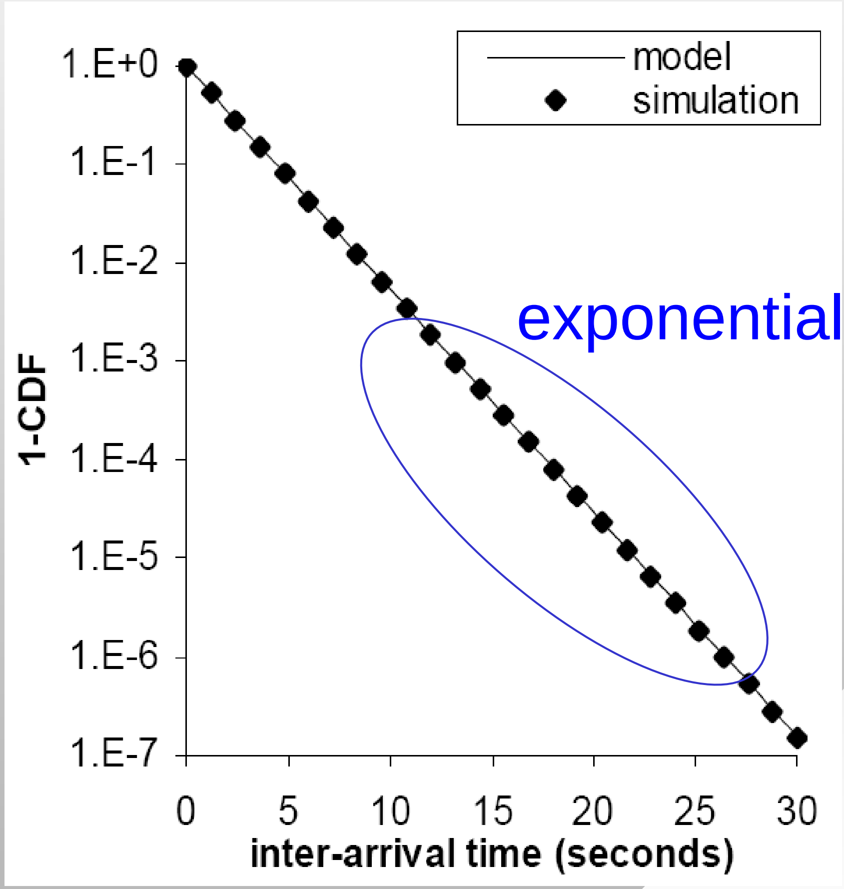
individual user arrival rate

- This Poisson result on user arrival in P2P networks is a consequence of our churn model rather than an assumption as in previous work
  - It does, however, show that prior assumptions on Poisson user arrival are valid approximations

User Arrival Process



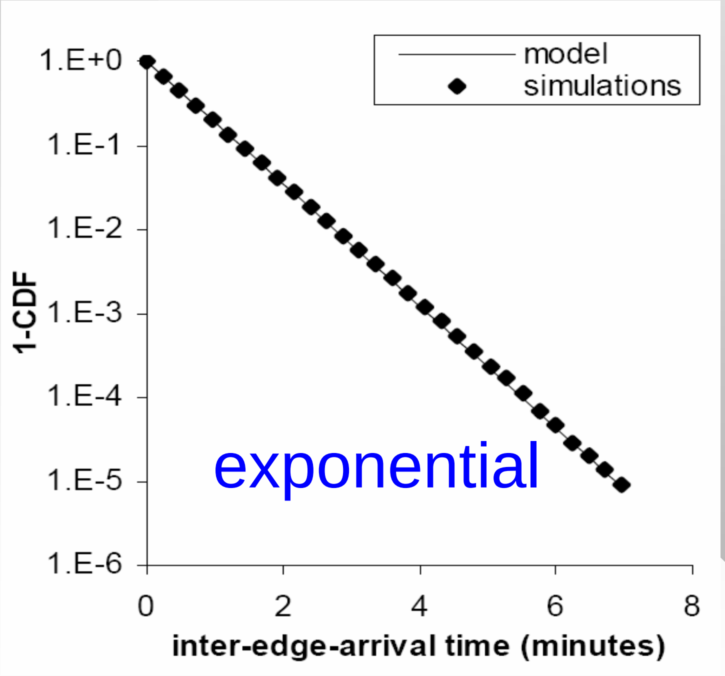
(a) $F_i(x)$ is heavy-tailed, $n = 20$



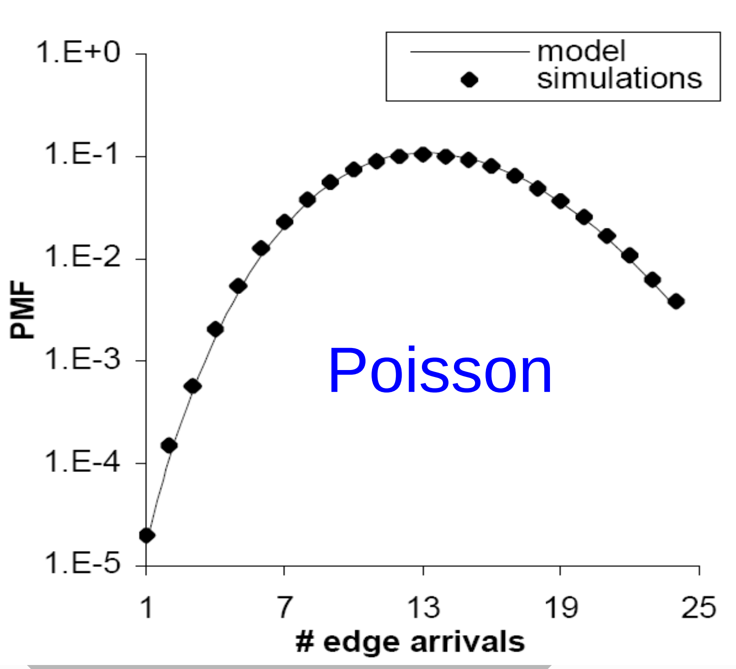
(b) $F_i(x)$ is heavy-tailed, $n = 1000$

Edge Arrival Process

- Theorem 7: Edge arrival to a random user v under uniform selection converges as $n \to \infty$ to a homogeneous Poisson process



(a) $F_i(x)$ is heavy-tailed, $n = 5000$



(b) time interval $\Delta t = 9$ min

In-Degree Model

- Let $X(t)$ be the random in-degree of a user v with current age $t \geq 0$
- Theorem 8: Under uniform selection, mean in-degree at age t is a monotonically increasing function of age t given by:

user lifetime distribution

$$E[X(t)] = \int_0^t \frac{k(1-(F(t-z)) + \theta(1-H_U(t-z))}{E[L]} dz$$

out-degree

in-degree at departure

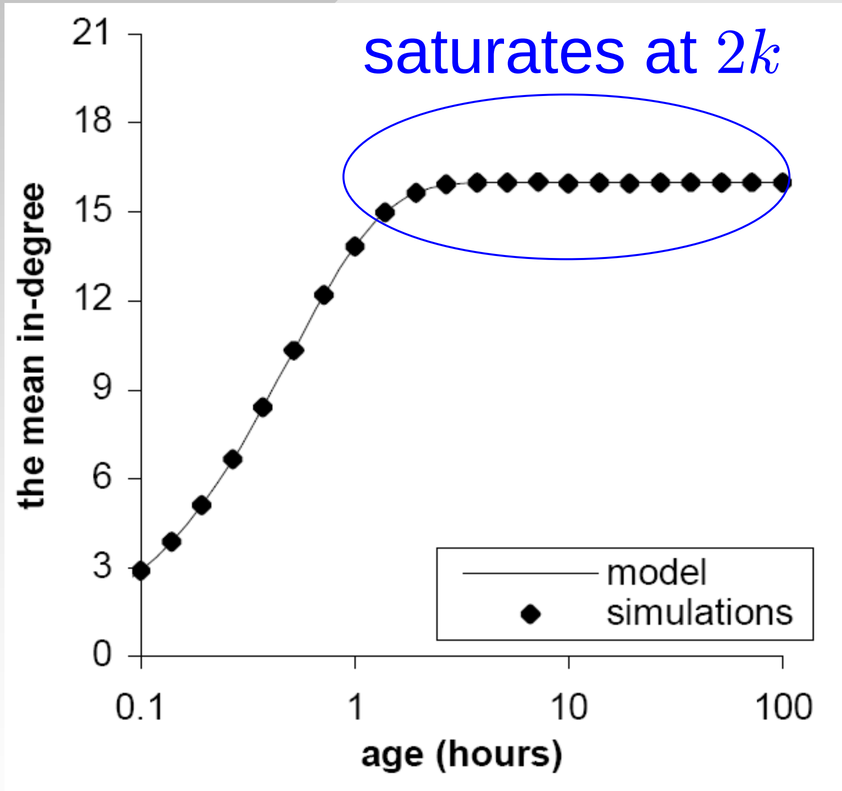
residual lifetime distribution

Moreover, $X(t)$ tends to a Poisson random variable

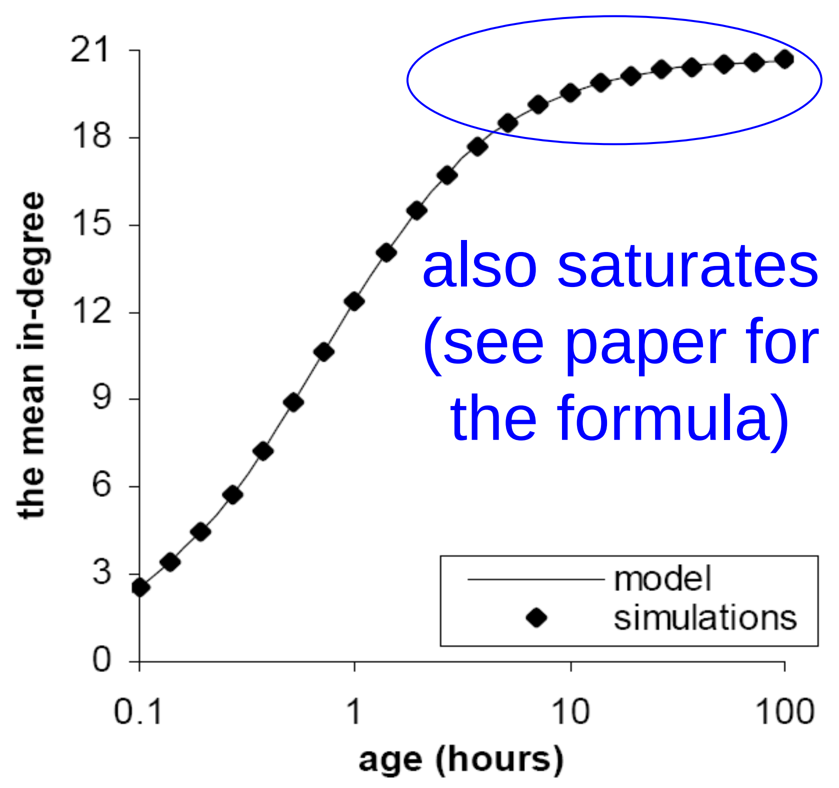
- Additional details and derivations in the paper

Expected In-degree

- Simulation results under uniform selection



(a) exponential lifetimes with $\mu = 2$



(b) Pareto lifetimes with $\alpha = 3$

Agenda

- Motivation and background
  - Terminology, assumptions, and previous work
- Heterogeneous churn model
  - Lifetime distribution of joining users
  - Residual lifetime distribution
  - Lifetime distribution of users in the system
- In-degree results (summary)
- Joint in/out-degree results (summary)
- Wrap up

Joint In/Out-degree Model

- Theorem 9: For exponential lifetimes $L \sim \exp(\mu)$ and exponential search delays $S \sim \exp(\sigma)$, node isolation probability converges to the following as $E[S] \to 0$:

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

out-degree isolation probability (Leonard 2005)

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

- Reduction in isolation probability by roughly a factor of $2k$ for non-trial k
  - Short-lived users do not benefit much; however, long-lived peers obtain significant benefit from the in-degree process, which leads to improved resilience of the entire system
- Refer to the paper for more discussion

Wrap-up

- We introduced a heterogeneous user churn model
  - Approximates user participation except two cases: dependence between lifetimes of different users and presence of each user under multiple identities
- Under uniform selection, we showed that the lifetime distribution of joining users was sufficient to completely model the effect of churn on P2P graphs
  - For these cases, we obtained closed-form results on the behavior of in-degree as a function of user age
  - We also derived the in/out-degree isolation probability and showed that users with large lifetimes significantly improved their resilience from the in-degree process