Analysis of wide area user mobility patterns

Kevin Simler*, Steven E. Czerwinski†, Anthony Joseph

UC Berkeley

2004/12/02

* Now at MIT

† Now at Google

Motivation

- We want to understand user behavior
  - In order to design better systems
  - In order to generate synthetic traces
  - In order to model user behavior
- How can we capture user presence in the wide area?

Motivation

- We want to understand user behavior
  - In order to design better systems
  - In order to generate synthetic traces
  - In order to model user behavior
- How can we capture user presence in the wide area?

web

Motivation

- We want to understand user behavior
  - In order to design better systems
  - In order to generate synthetic traces
  - In order to model user behavior
- How can we capture user presence in the wide area?

web, IM

Motivation

- We want to understand user behavior
  - In order to design better systems
  - In order to generate synthetic traces
  - In order to model user behavior
- How can we capture user presence in the wide area?

web, IM, …, e-mail

Why e-mail?

- E-mail is a widely-used service
- User typically checks e-mail first
- Berkeley provides IMAP + web front end
  - Any Internet connection → e-mail access

- E-mail reflects users' Internet presence

Outline

- Background
- Analysis and results
- User modeling
- Future work
- Summary

Trace characteristics

- 31-days (May 2003)
- Server from UC Berkeley EECS dept.
  - Regular IMAP plus web front-end
- 1004 active users, primarily:
  - Professors
  - Graduate students
  - Support staff
- Tracked across different service providers

Building on previous work

- Wireless Campus Studies
  - Mobility on a campus
  - Single service provider with homogenous users
  - Tang & Baker, Kotz & Essien, Balazinska & Castro
- Metricom WLAN
  - Mobility across/between cities
  - Single service provider with more diverse users
  - Tang & Baker

Trace data

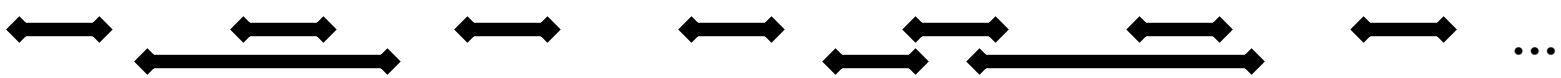
- Each entry in the trace includes:
  - Timestamp (seconds)
  - Request type (*login*, *close*, *select*, etc.)
  - Username
  - IP address

Preprocessing

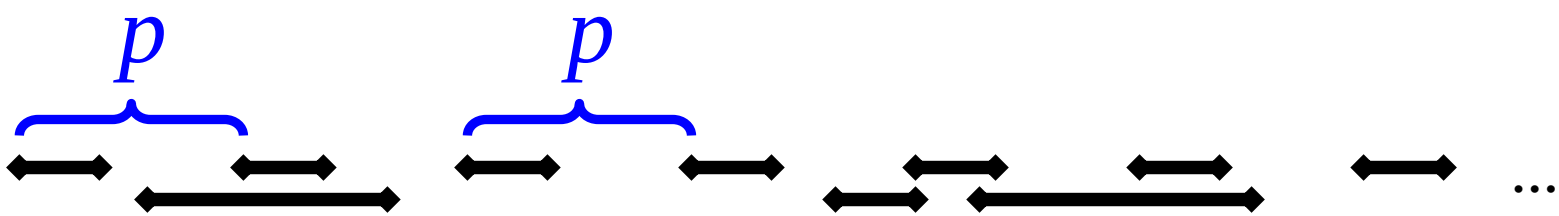
- We want **user behavior**
- Trace records **client application behavior**
  - Outlook, Eudora, Thunderbird, etc.
- Primary difference:
  - Client polls for new e-mail at regular intervals
  - Fixed period per client, variable across clients

We filter client polling using a Fourier transform

Client connections from a single user:



We filter client polling using a Fourier transform



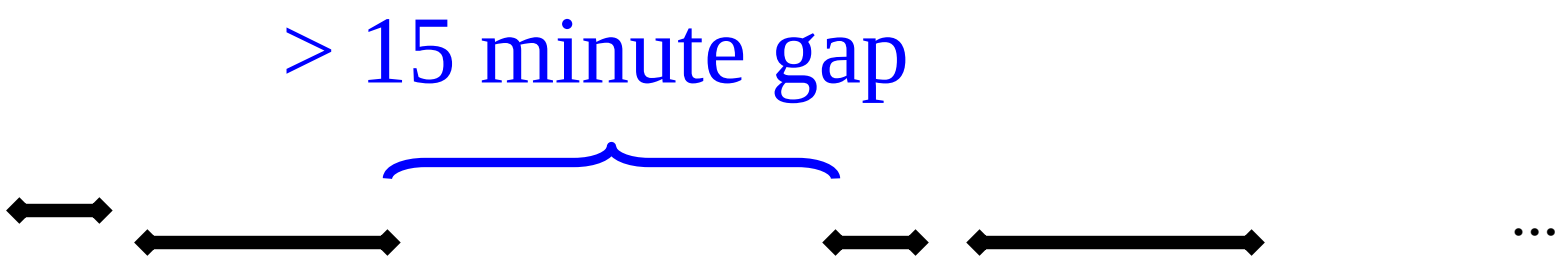
Use a Fourier transform to identify polling period *p*.

We filter client polling using a Fourier transform

Identify sequence separated by *p*.

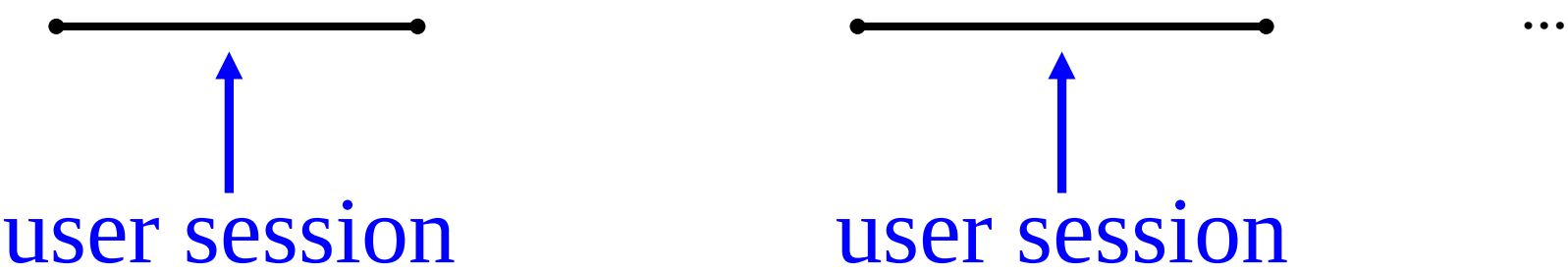
Remove all but the first connection.

We filter client polling using a Fourier transform



Clump connections into user sessions

We filter client polling using a Fourier transform



We filter client polling using a Fourier transform

Now we have (roughly) a trace of user behavior

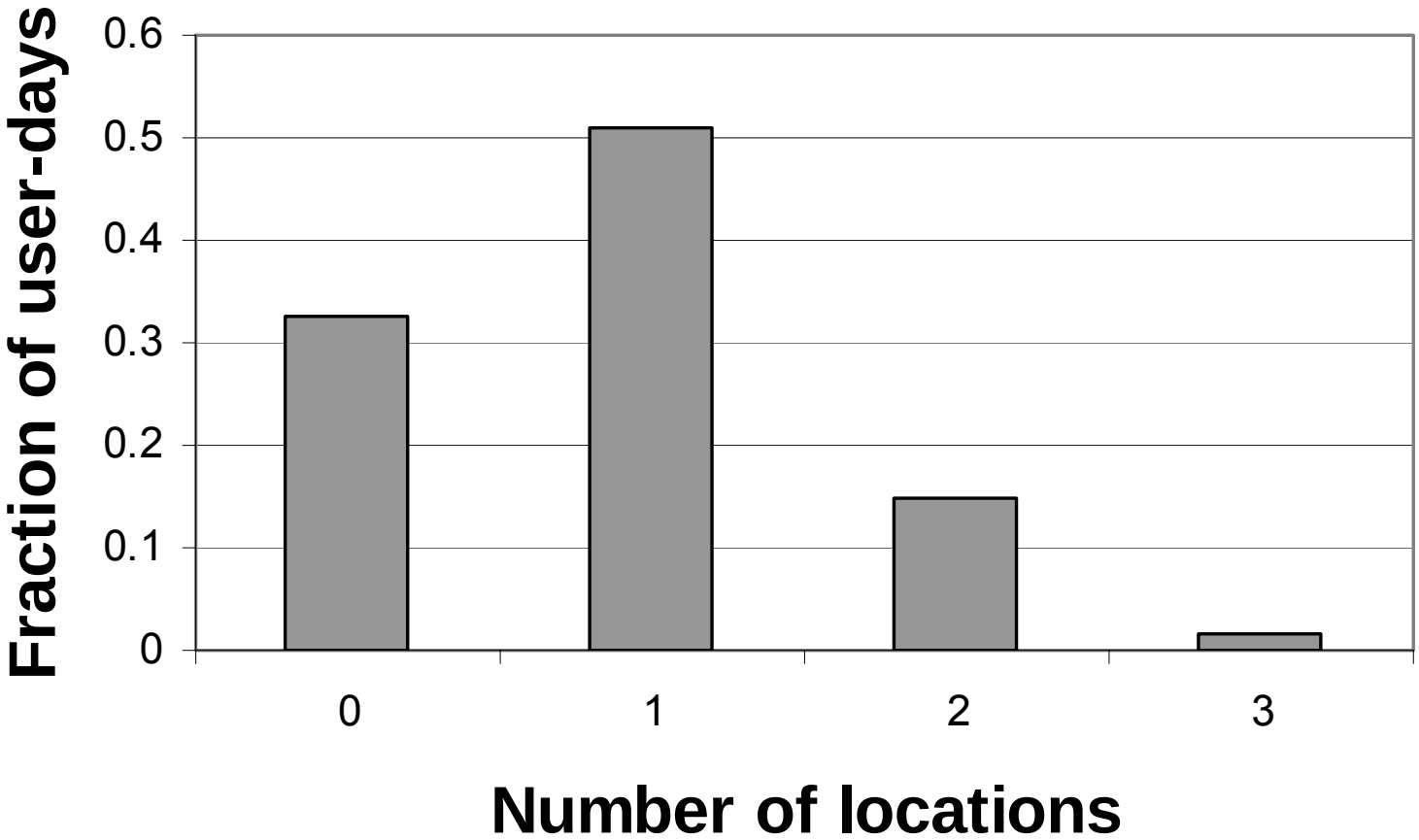
Outline

- Background
- **Trace analysis**
  - Defining location
  - Daily mobility
  - Monthly mobility
  - Session activity
- User modeling
- Future work
- Summary

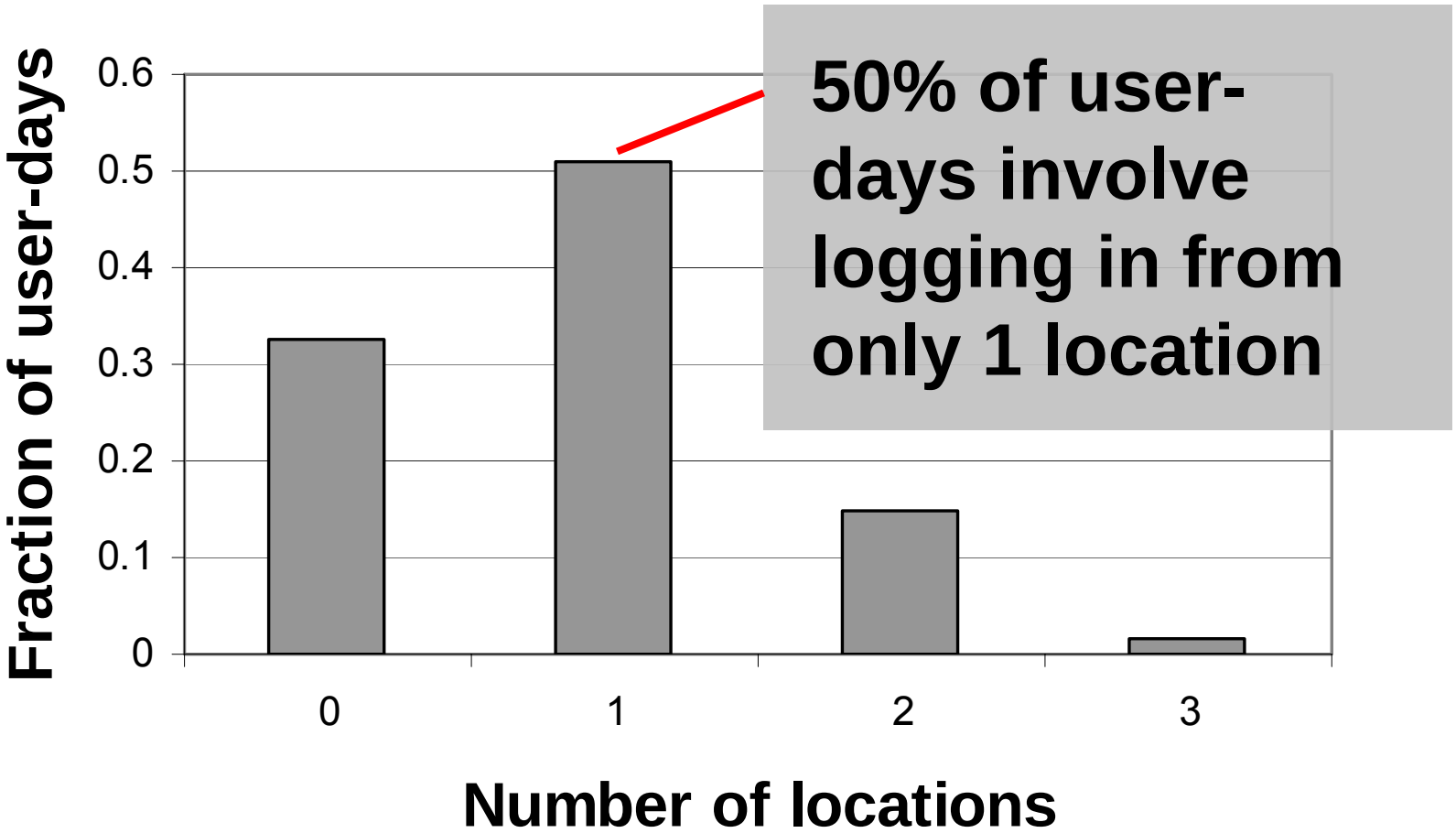
Defining network location

- Connection used to access the Internet
  - E.g. a dialup ISP, campus wireless network
- Approximated by a combination of
  - Authoritative DNS server
  - AS number
  - Subnet

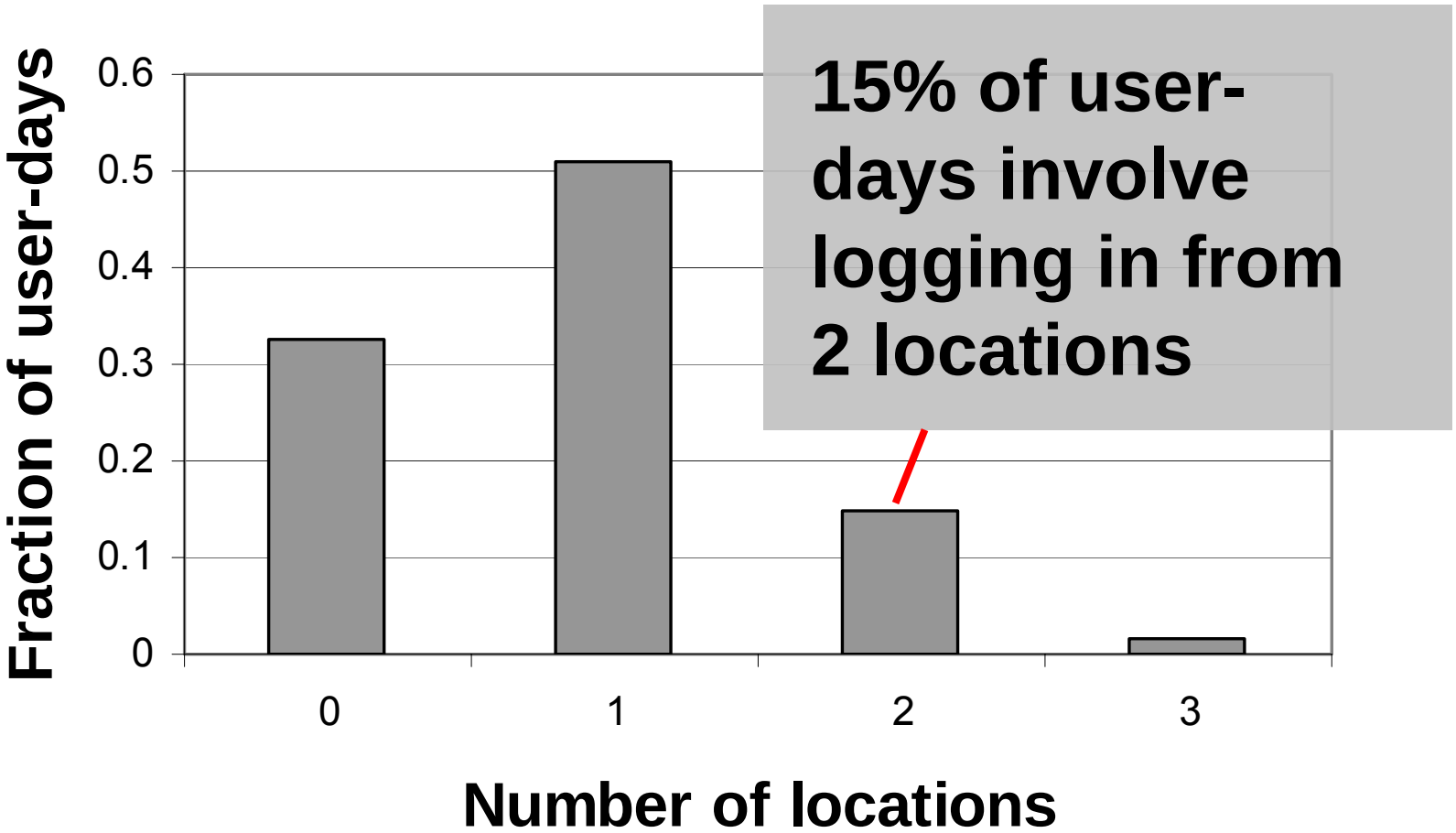
How mobile are users each day?



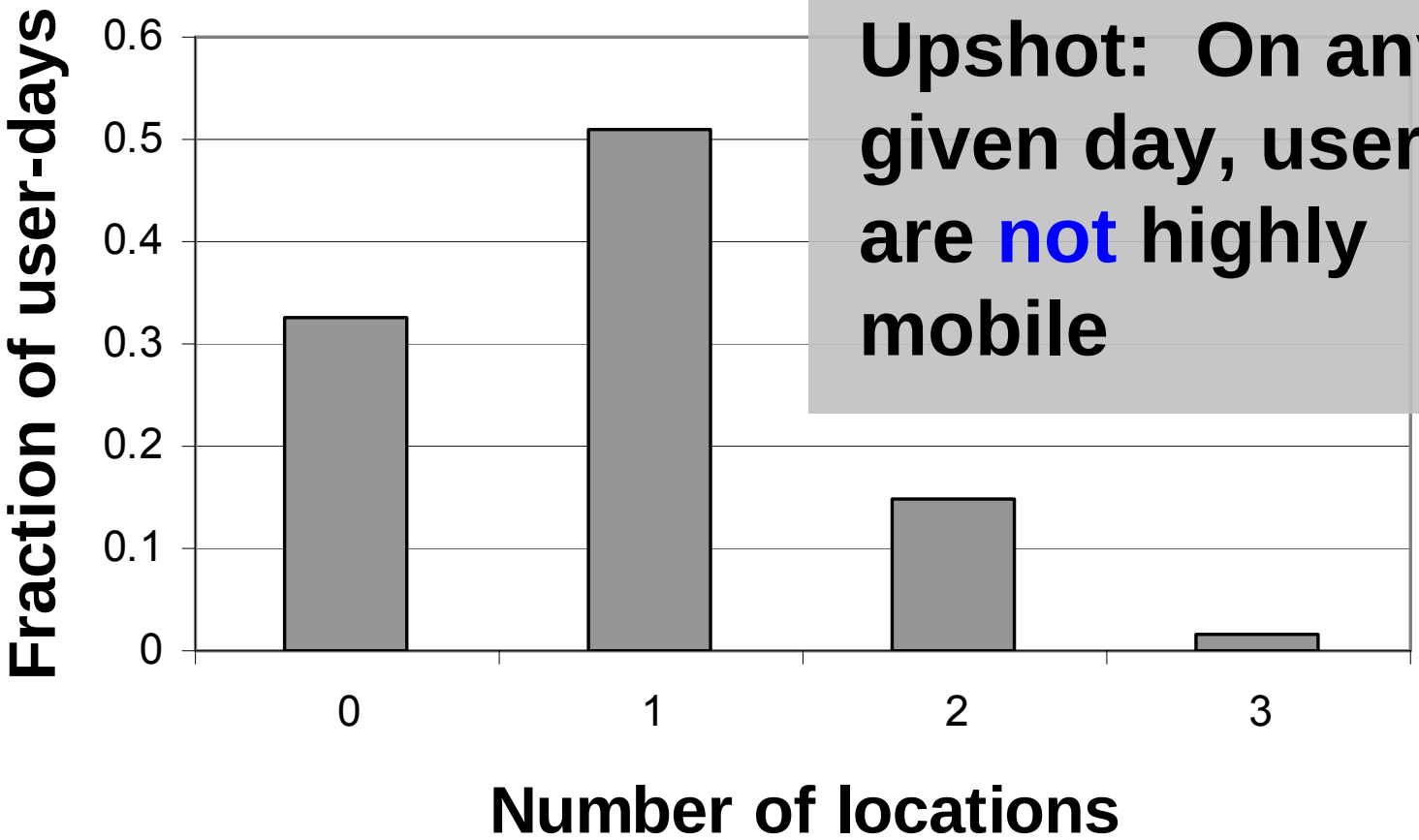
How mobile are users each day?



How mobile are users each day?



How mobile are users each day?



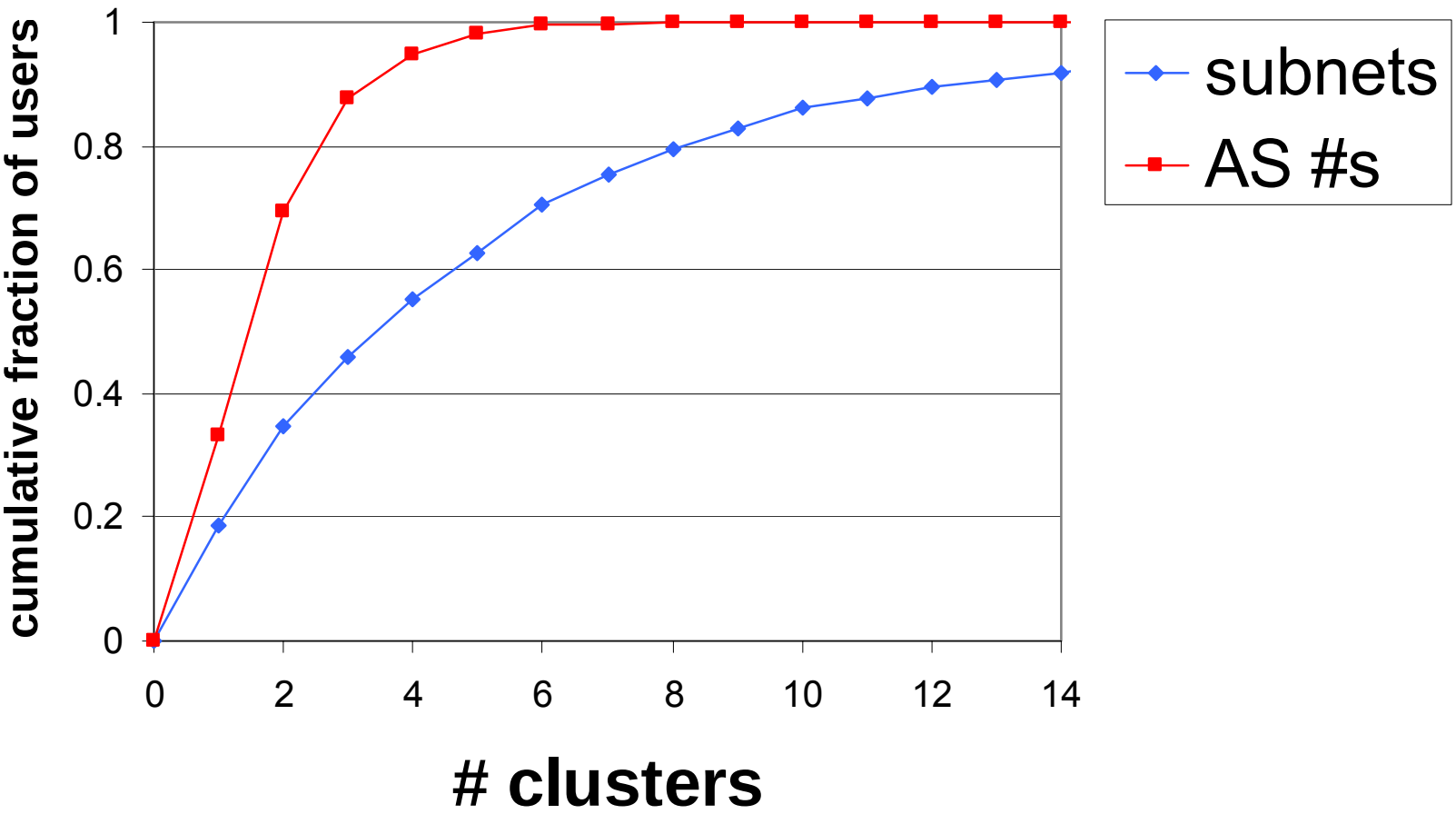
Upshot: On any given day, users are not highly mobile

How mobile are users in 31 days?

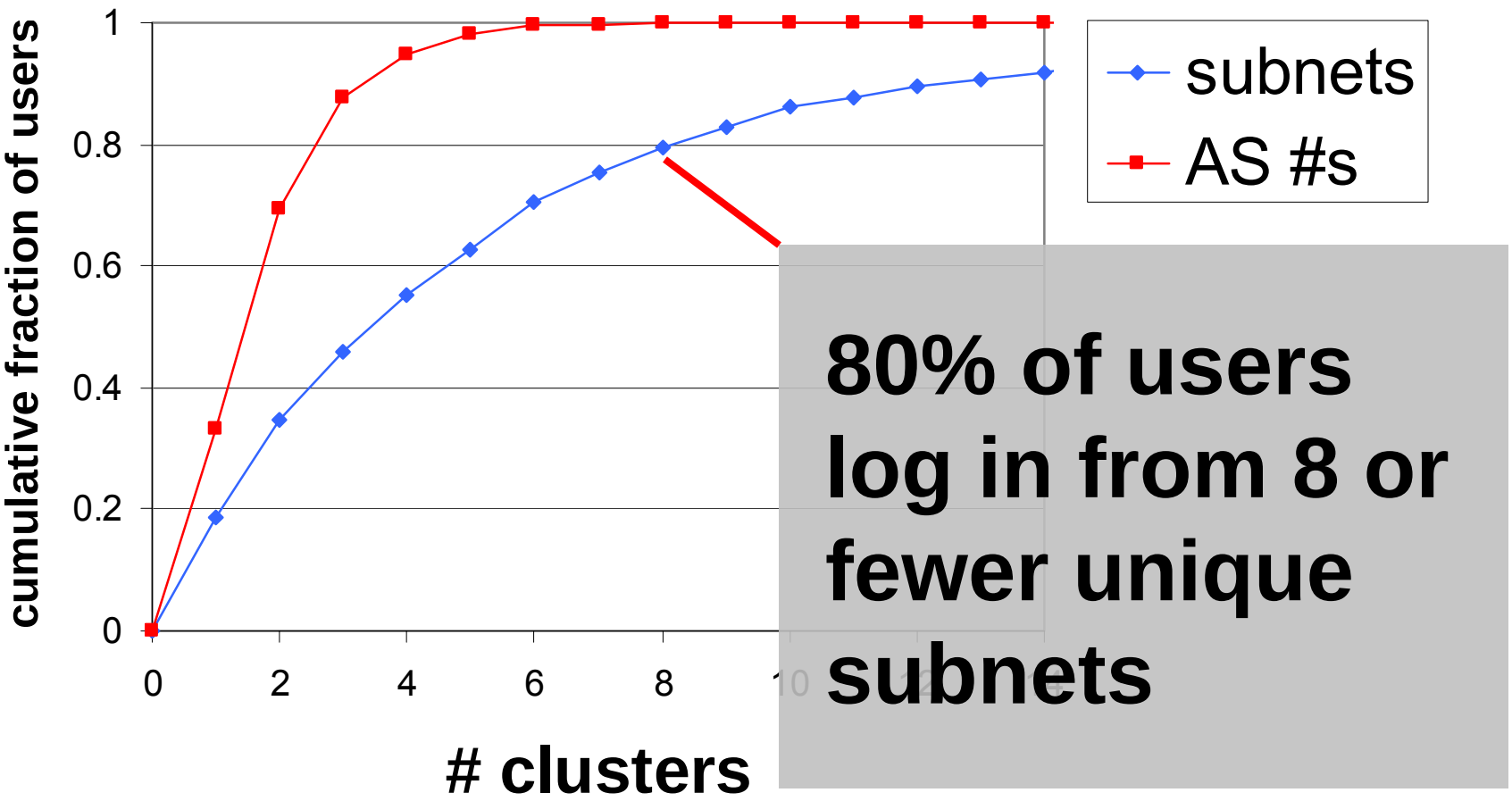
- How many unique subnets do they visit?
- How many unique AS #s do they visit?

Let's look at a graph….

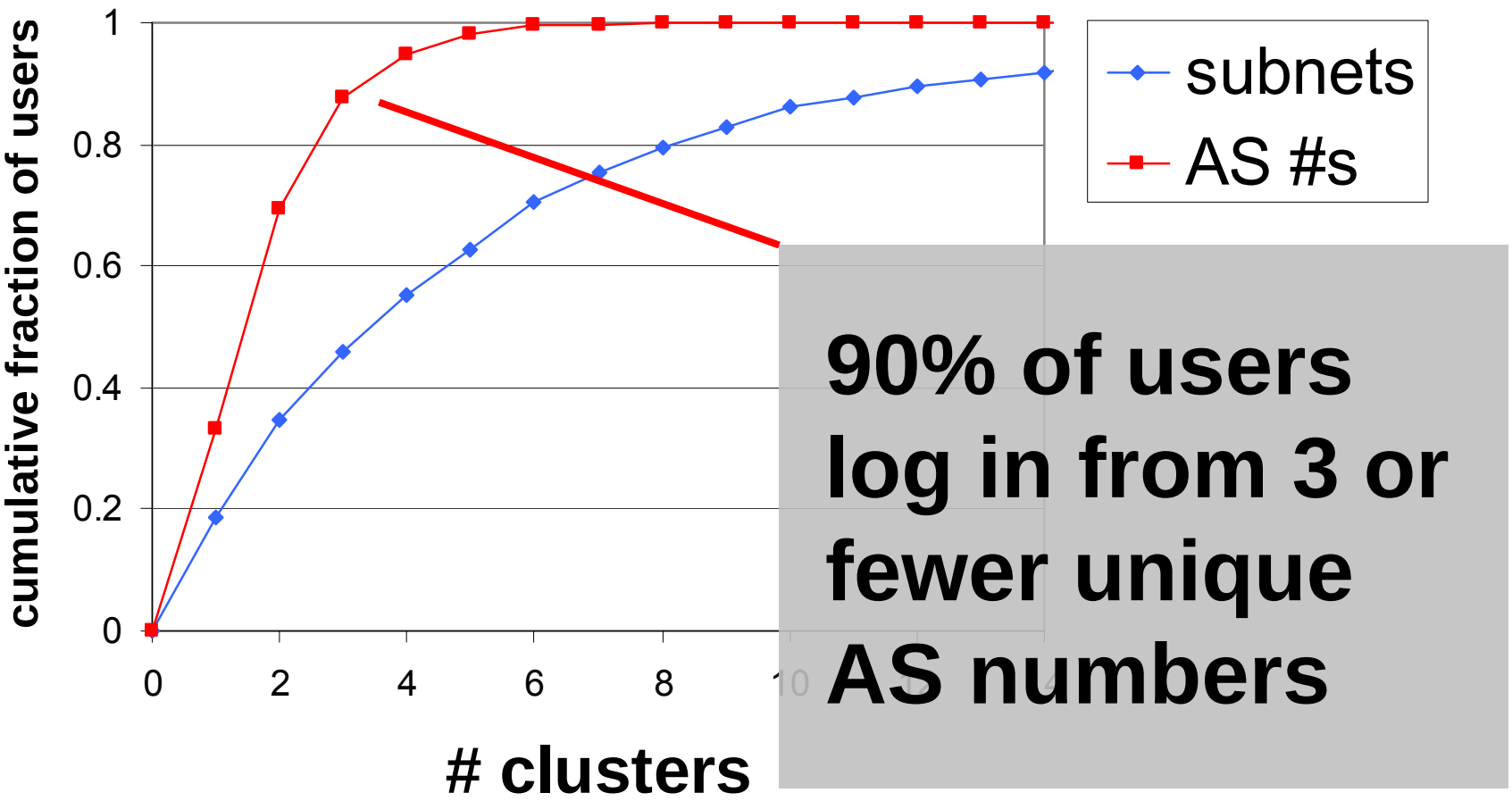
How mobile are users in 31 days?



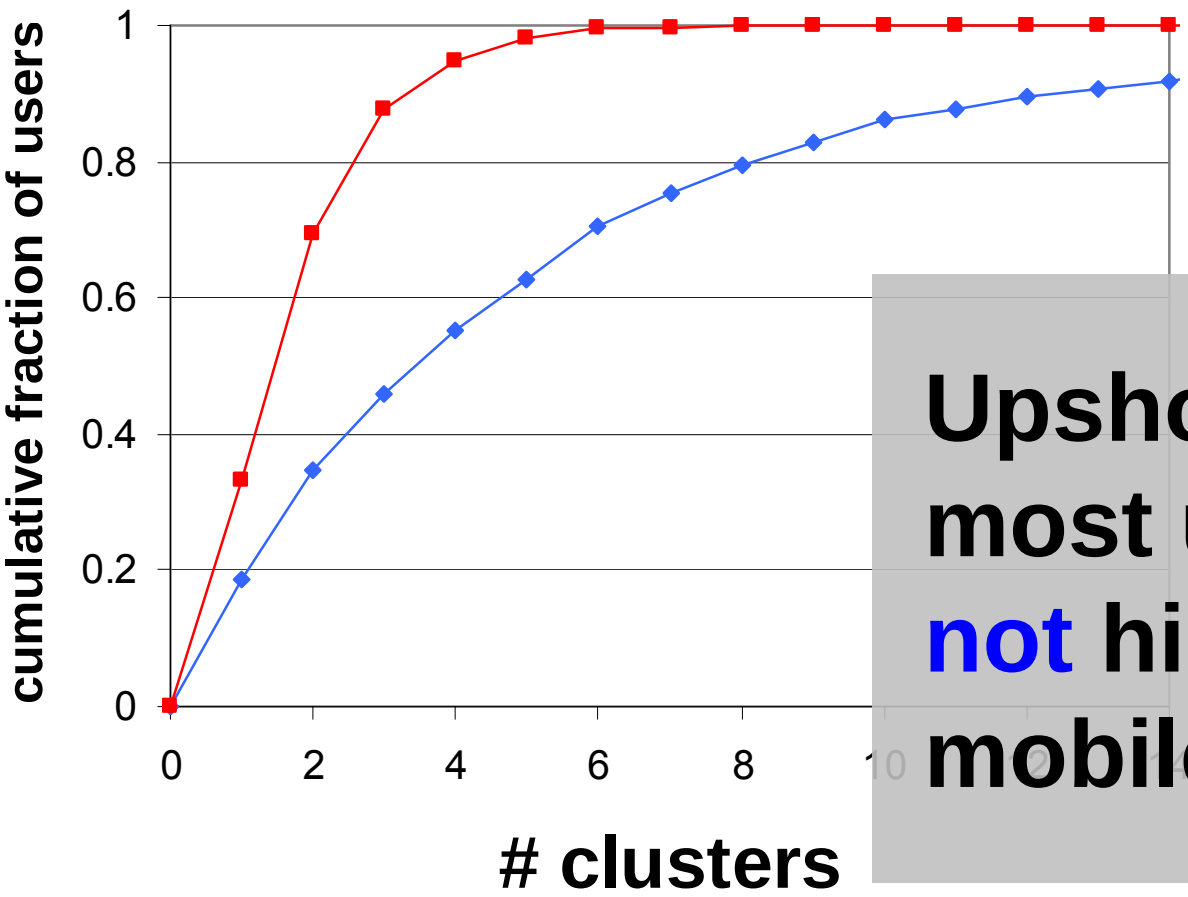
How mobile are users in 31 days?



How mobile are users in 31 days?

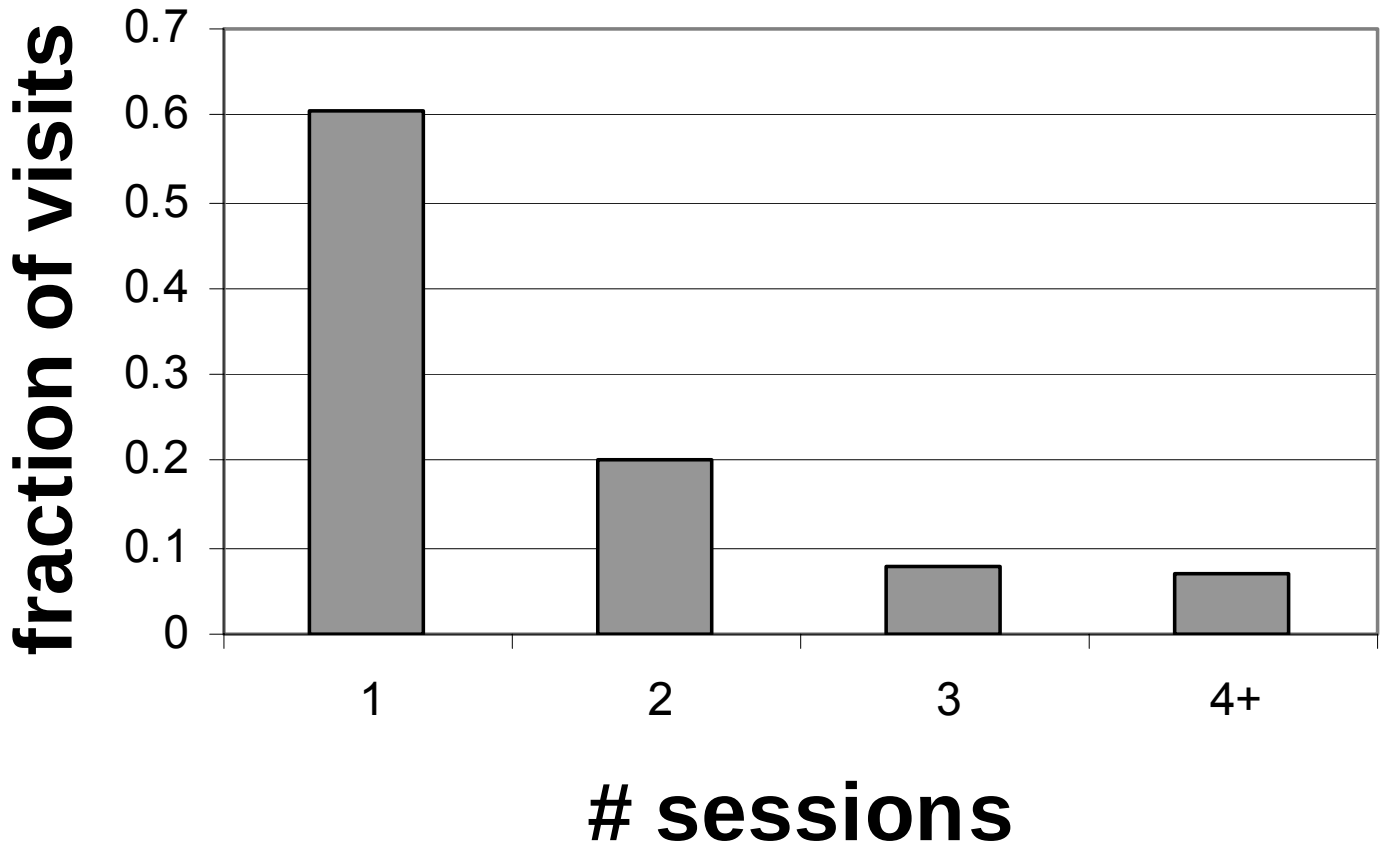


How mobile are users in 31 days?

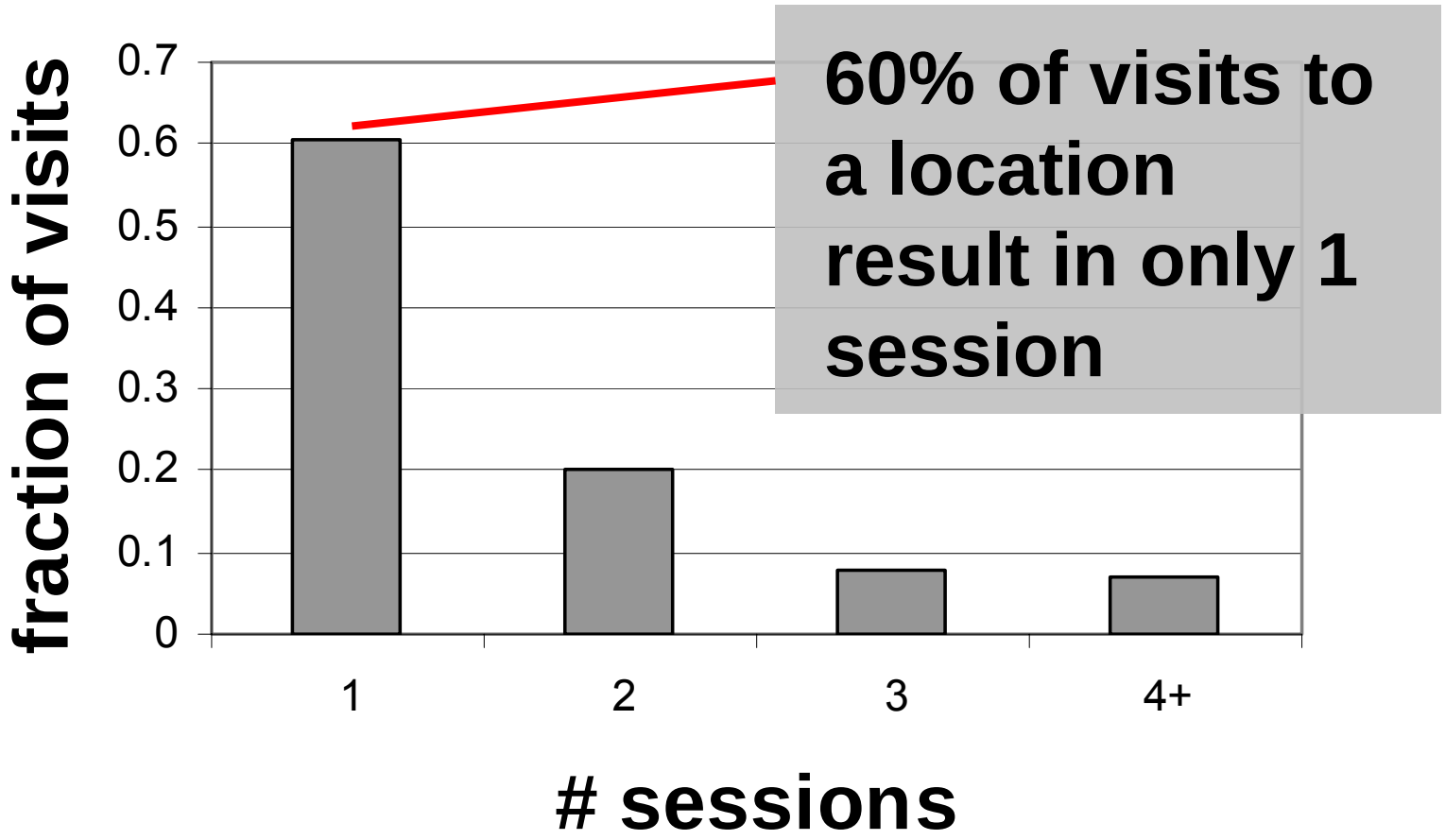


Upshot: Again, most users are not highly mobile

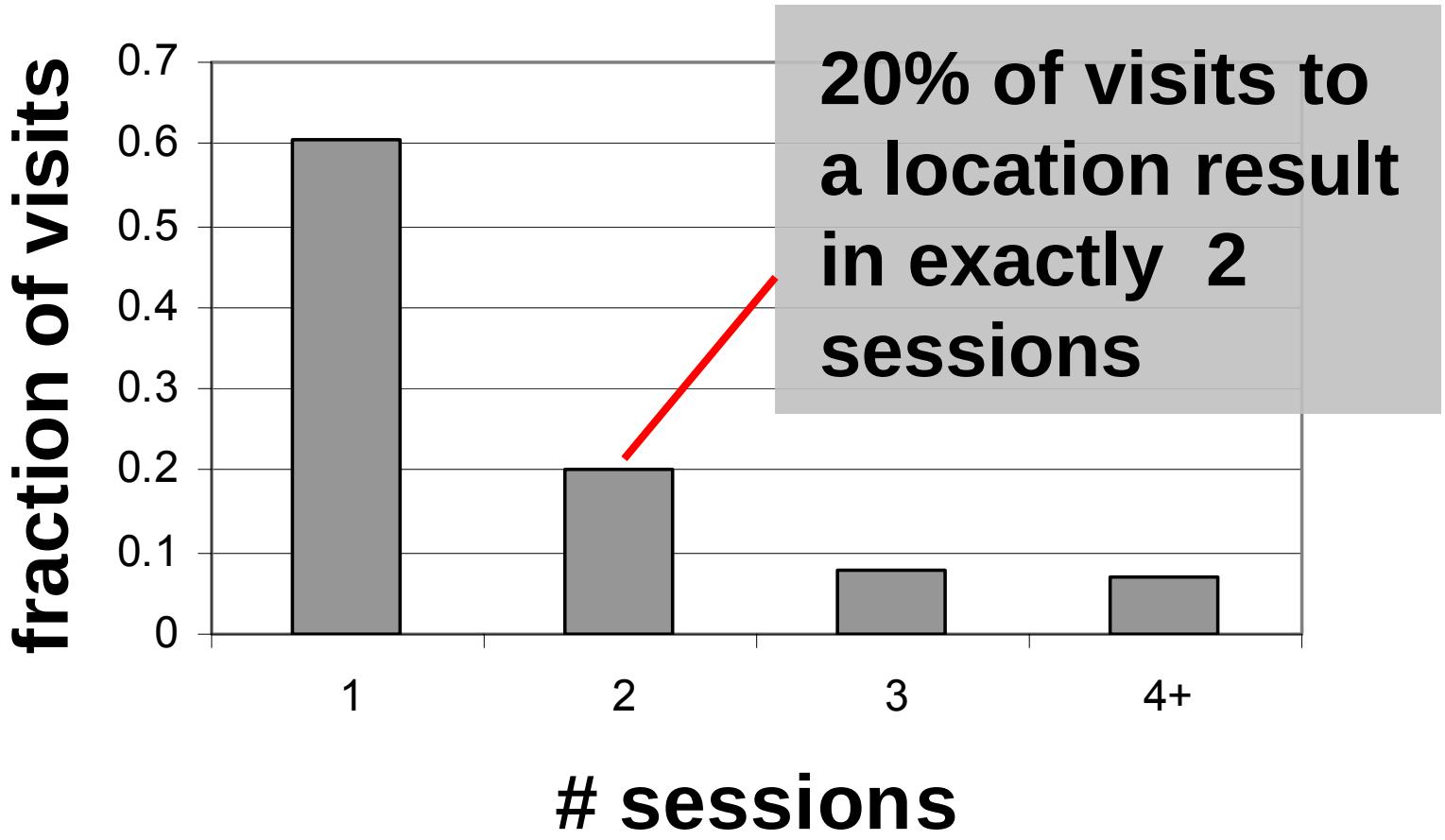
User activity at a location



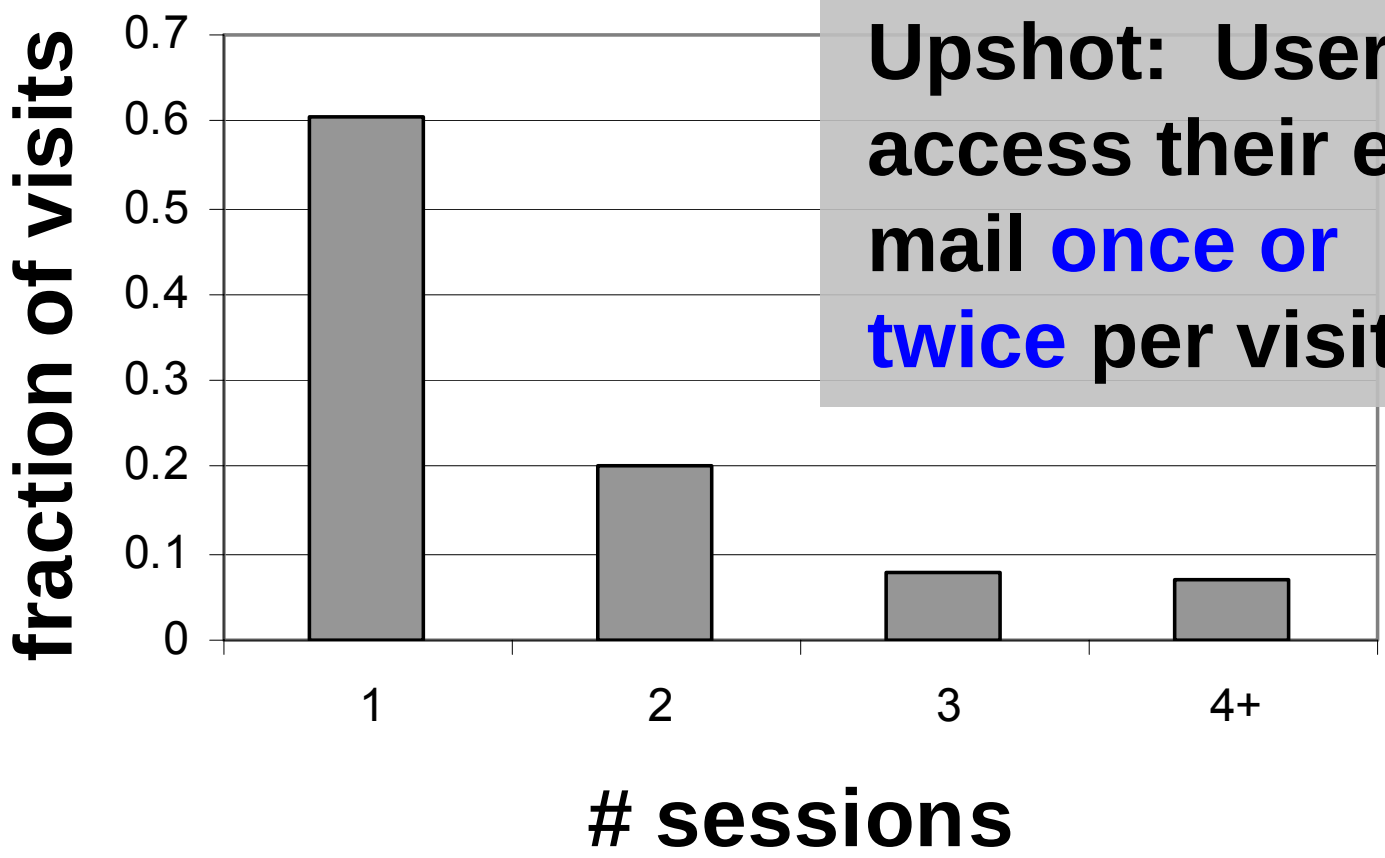
User activity at a location



User activity at a location



User activity at a location



Upshot: Users access their e-mail once or twice per visit.

Outline

- Background
- Trace analysis
- **User modeling**
  - Categorizing users
  - Model structure
  - Training and testing
- Future work
- Summary

Categorizing users

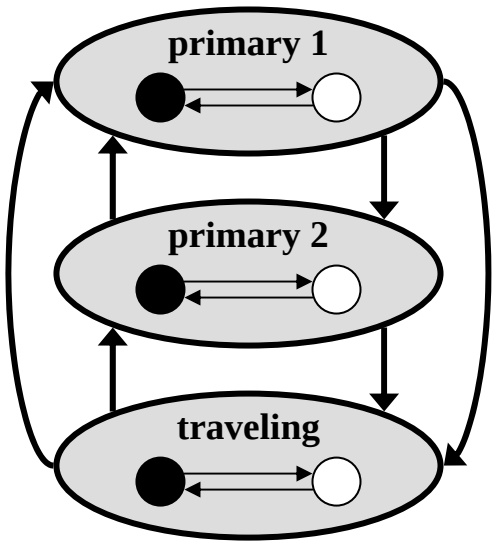
- Based on number of primary locations
- For a given user, a primary location is:
  - One where the user spends >5% of the time
- Categories
  - Users with 1 primary location
  - Users with 2 primary locations
  - Users with 3+ primary locations

Structure of our models

- One model for each category
- Two-tiered Markov model
  - High-level states represent user's location
  - Low-level states represent user's activity
- Both MMs are 1st order

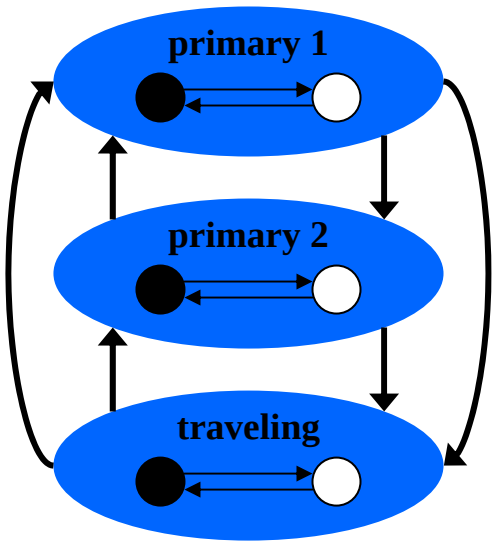
Model structure for category 2

- 2 primary locations + 1 traveling state



Model structure for category 2

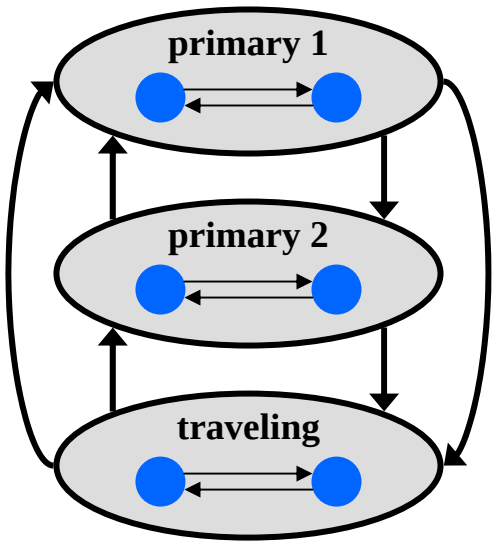
- 2 primary locations + 1 traveling state



High-level (location) states

Model structure for category 2

- 2 primary locations + 1 traveling state



Low-level (session) states

I.e. Logged-In and Logged-Out

Training

- We have all the information
  - Which locations are primary
  - Where the user is, at any time
  - When the user is logged in/out
- Simple to compute transition probabilities

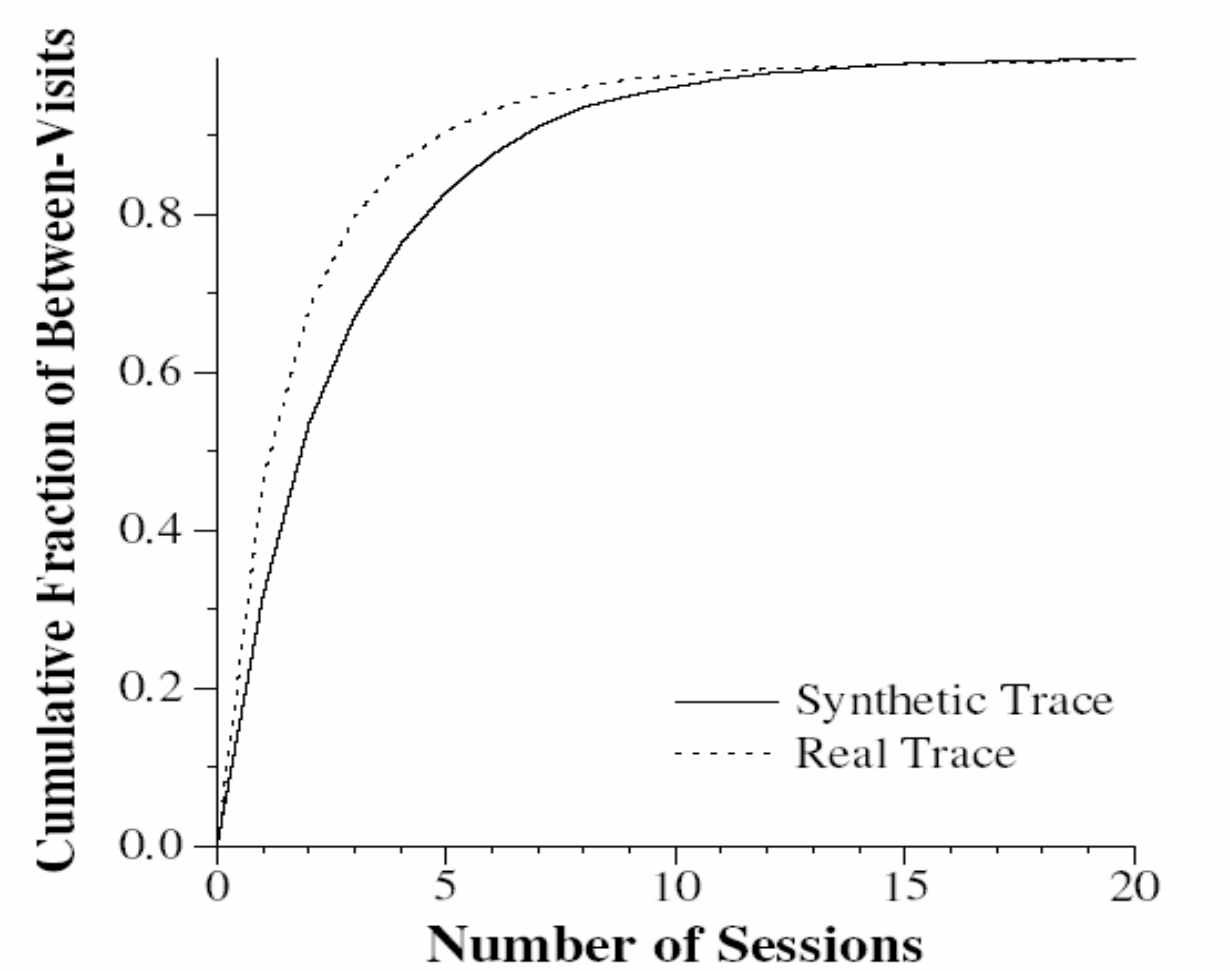
Testing methodology

- Create synthetic trace
- Chose metrics to measure a trace
- Compare real trace with synthetic trace

Testing one metric

- # of sessions between visits to primary
  - Each user visits his primary
  - leaves to visit other locations
  - then comes back to his primary
- Every time this happens, record the number of other locations
- There will be a CDF for the entire trace (real or synthetic)

Testing results



Outline

- Background
- Trace analysis
- User modeling
- **Future work**
- Summary

Using the results

- Synthetic traces can help test systems
- User behavior has implications for design
  - E.g. focus resources on primary locations
- Model can predict user behavior on-the-fly
  - E.g. to cache, or not to cache?

As technology changes…

- Blackberries
  - More physical locations
  - Shorter, more frequent sessions
  - Still, primary locations will be important

- Wireless LAN hotspots
  - More network locations

Outline

- Background
- Trace analysis
- User modeling
- Future work
- **Summary**

Summary – what we've done

- Obtained a trace from an e-mail server
- Filtered out client polling
- Analyzed trace of user behavior
- Modeled categories of users with tiered MM
- Generated synthetic traces

Summary – user behavior

- Most users log in from 1 or 2 locations
- But a few users are highly mobile
- Users access e-mail infrequently, but for long periods of time

Thank you

- Quick clarifying questions?