Methods for Creating Metadata for Shareable Files

Carlos Maltzahn
Assistant Adjunct Professor
University of California, Santa Cruz

and
Nikhil Bobb,
Mark W. Storer,
Damian Eads,
Scott A. Brandt,
Ethan L. Miller

Problem

- My laptop:
  - 144GB consisting of
  - 1.3m files (23k, ~2% under Documents) in
  - 300k directories
- 24,738 registered file types
  [filext.com -- 6/25/08]
- So what?
  - Mounting data loss
  - Digital age becomes a dark age

Common Approaches

- Directories
  - hard to create and remember categories [Lansdale'88]
- File system-wide Search
  - insufficient expressive power [Lansdale'88]
  - scarce metadata: content insufficient, little context
  - high expectations for accuracy
- Information Mgmt Applications
  - metadata silos, no infrastructure/standards
  - file type-specific
  - no infrastructure/standards to share

Collaborative Tagging?

- Web has many successful examples:

Delicious Digg reddit Facebook StumbleUpon

- 28% of Internet users [Pew'06]
  - Complemental to web search

- Enables recommender systems
  - Eases cognitive load of assigning and remembering tags

Questions

- Does the success of collaborative tagging transfer to file systems?
  - Are there enough shareable files?
  - How to name shared files?
  - How about structures within files?
  - How to share metadata of files?

- How do people find time for this?

Shareability Hypothesis

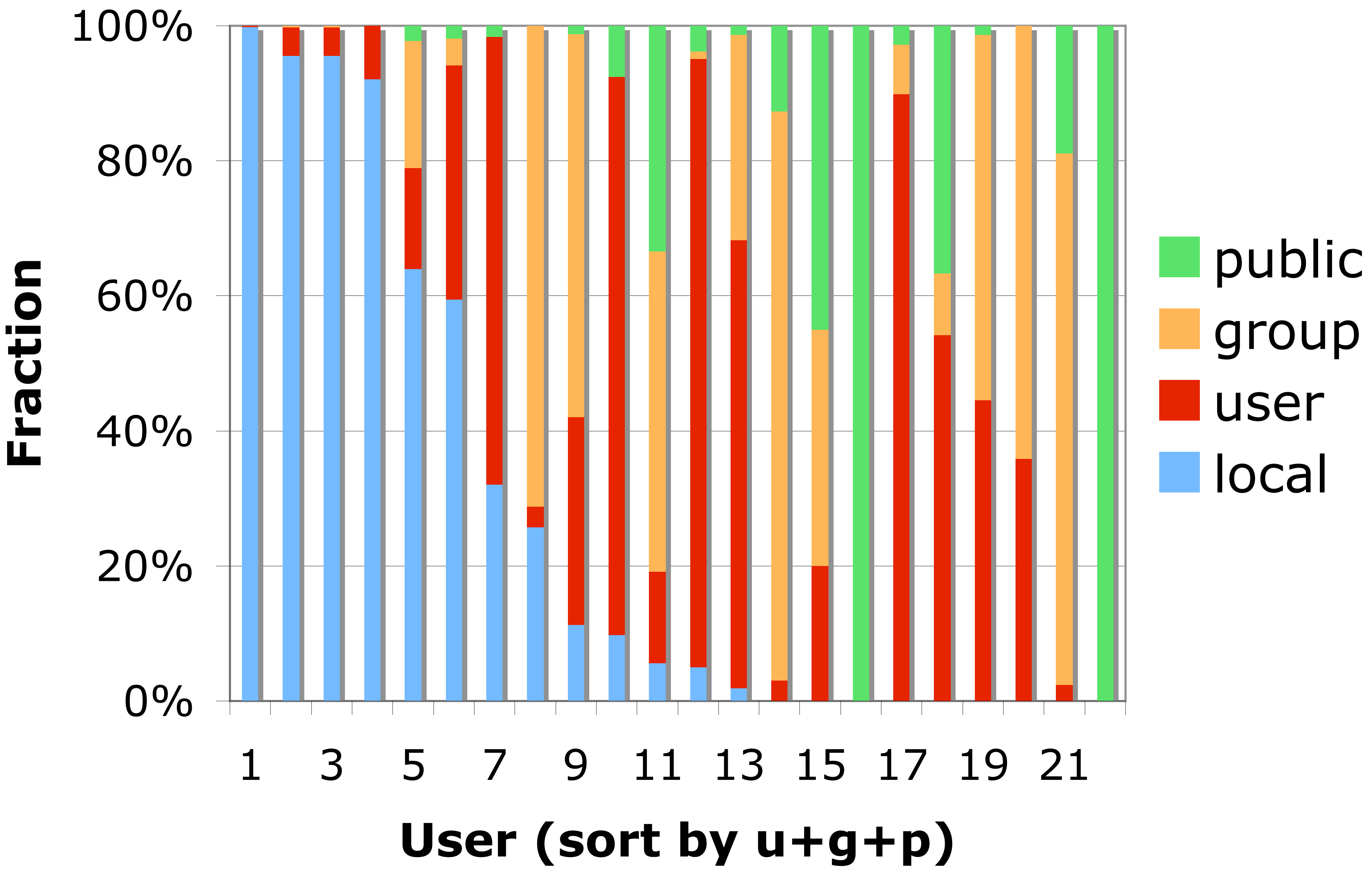
- *shareable file*:
  - *should be* managed across file systems and/or users.
  - not necessarily managed that way, currently.

File systems have enough *shareable files* to make collaborative data management feasible.

Categories

- **skipped:** file is unknown, uninteresting, or irrelevant
- **local:** file never leaves this computer (not shareable)
  - user wants to manage file
  - not suitable for sharing among computers or users
- **user:** file is private (shareable)
  - suitable for sharing among computers of same user
- **group:** file is restricted to group (shareable)
- **public:** file is public (shareable)
  - downloaded or published files

Results:

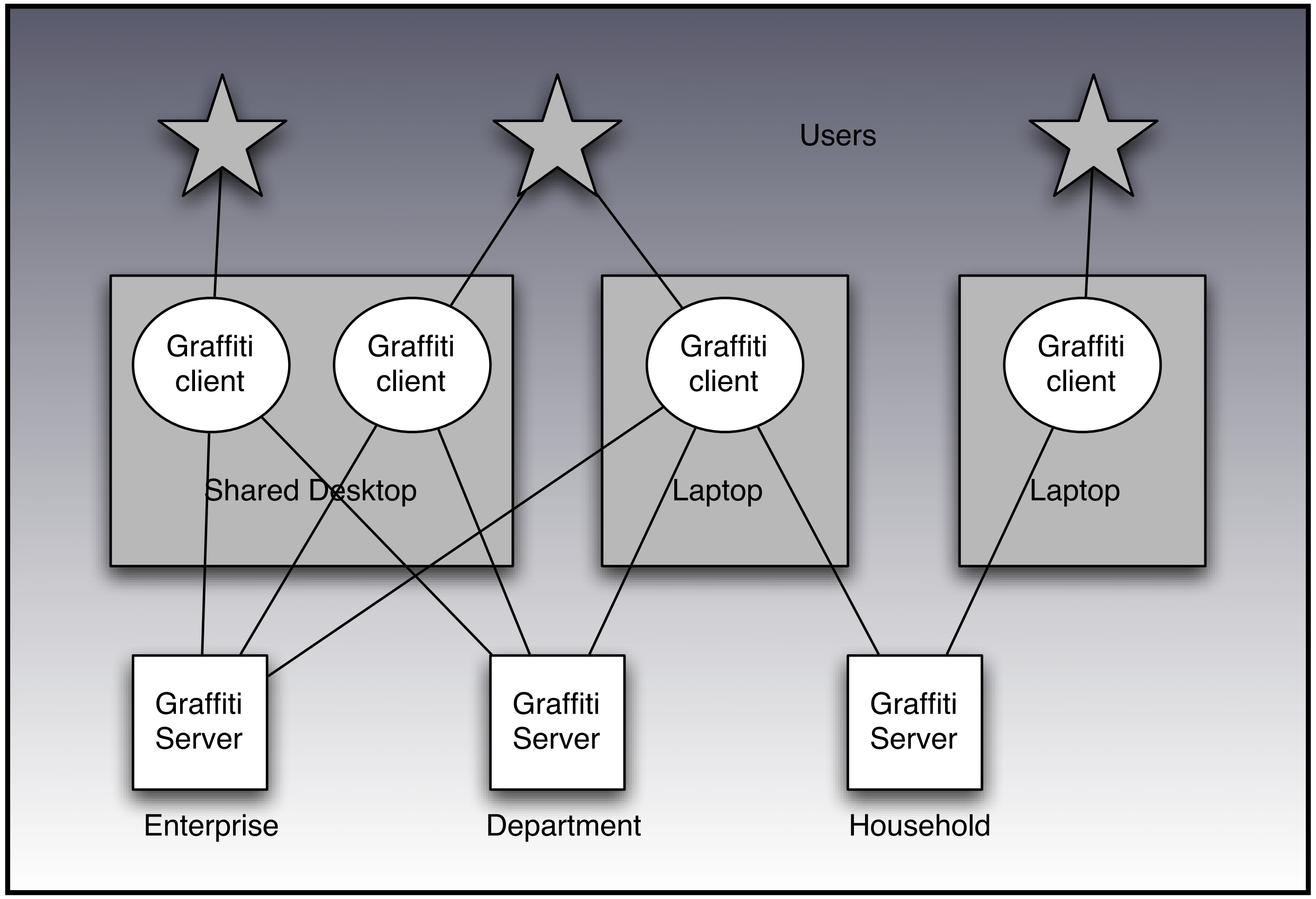


85% of surveyed & relevant files are shareable!

Reactions

- Multiple respondents didn’t take the survey: little or no concept of “files” or “home directory”
- “If I want to share files I move them to our [home/public] server”
- “Are my tax returns ‘private’ or ‘group’ because I eventually share them with the IRS?”
- “You forgot the ‘enterprise’ category”

Graffiti



- Think IMAP email accounts for file metadata
- Users control metadata proliferation to specific servers
- Clients maintain local metadata and coherence with servers

Content-based File Naming

- Global, file system-independent names
- No central naming authority
  - Discovery
  - Scalability
- Checksum overhead can be minimized by landmark chunking, etc.
- Provides similarity metric for recommenders

Experience

- Hierarchical directories are **not** obsolete
  - directories are good for hiding
  - tags are good for finding
- Tagging directories is useful
- Uncovering duplication has great potential
- Tagging, searching, and file browsing are interleaved

Do people have time for this?

Do people have time for this?

- Fundamental problem:
  - Information management demands increasing amount of attention
  - Where is the time for this?

Do people have time for this?

- Fundamental problem:
  - Information management demands increasing amount of attention
  - Where is the time for this?
- Approach:
  - Tapping into “Solitaire cycles”
  - Information management as a game

UCSC Computer Game Design

- Michael Mateas (tenured)
- Noah Wardrup-Fruin (just hired)
- Arnav Jhala (just hired)

Inspiring Examples

- Luis von Ahn (CMU):
  Human-based computation
  - Images: reCAPTCHA, ESP Game (Google Image Labeler), Peekaboom, Squigl, Phetch
  - Audio: Tag a Tune
  - Common Sense: Verbosity
- UW, Rosetta@home:
  - Foldit

Thank You !

carlosm@cs.ucsc.edu
www.cs.ucsc.edu/~carlosm

This work was supported in part by the Department of Energy under award DE-FC02-06ER25768.