How Shareable are Home Directories?

Carlos Maltzahn
Computer Science Department
University of California Santa Cruz
carlosm@cs.ucsc.edu

Problem

- People are increasingly overwhelmed by their data collections.
- Difficult to manage without search:
  - Large number of files
  - Large name space hierarchies
- Poor keyword search performance:
  - Scarce metadata (e.g. few relationships between files)
  - Small sets of relevant files (accuracy requires *more* metadata)

So what?

- People's file collection sizes increase exponentially
- Compounding effect of poor data management on cost of:
  - Data safety: more needs to be backed up
  - Data security: more needs to be encrypted
- Increased data loss due to cost limitations

Shareability Hypothesis

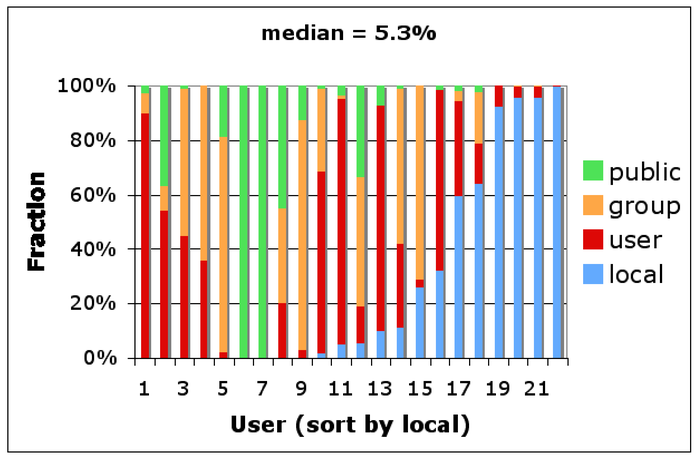
- Idea: Maximize sharing of data management effort
- Many successful examples on the Web:
  Delicious Digg reddit Facebook StumbleUpon
- 2006 Pew survey: 28% of Internet users have “tagged” online content
  - People spend time managing their files if payback sufficient

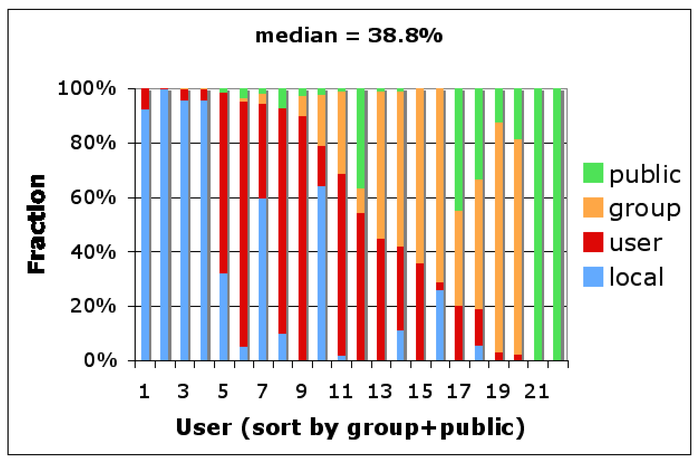
File systems have enough *shareable files* to make collaborative data management feasible.

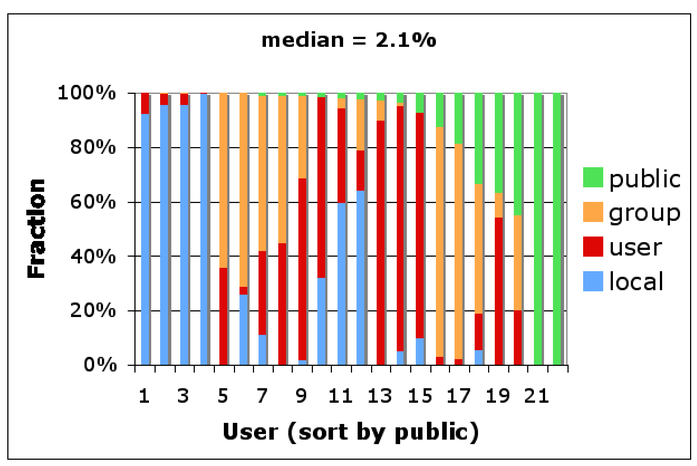
Potentially shared

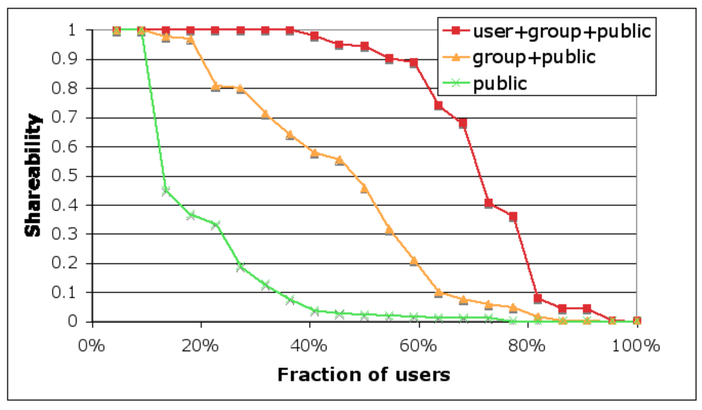
Approach

- *Shareable file*: managed across file systems and/or users
  - files that *should be* synced among home and work computer
  - files that *should be* shared with friends/colleagues
  - files that are downloaded from the web or shared publicly
- Here: Determine potential of shareable files in file systems
- Compare home directories among volunteer group
  - Over-estimate due to common systems and application files
  - Under-estimate due to sharing outside the group
- Instead: Subjective categorization
  - gage sharing potential
  - focus on files user cares about, skip the rest
  - focus on user-managed files
  - amount of sharing independent of sample size and available technologies









Categorization

- **local**: file never leaves this computer:
  - user wants to manage file
  - file not suitable for sharing among computers or users
- **user**: file is private:
  - file suitable for sharing among computer
  - file not suitable for sharing among different users
- **group**: file is restricted to a group:
  - file suitable for sharing among restricted group of users
- **public**: file is public
  - downloaded files from the web
  - published files

Survey

- Solicitation of colleagues, friends, and family
- Subject downloads small application (“ugo”):
  - walks through home directory hierarchy
  - speeds up categorization: single key stroke interaction, undo, redo, and entire directories with one key stroke
  - quit, resume: maintains state between sessions
  - produces result file
  - extra benefit: supports “trash” marking of files
  - reduces bias towards computer literacy
- Subject submits result file anonymously via web page
- UCSC IRB approved

Preliminary Results

- 75% of users show more than 50% shareability
  - user + group + public
- 50% of users show more than 50% shareability across users
  - group + public
- 10% of users show more than 50% public shareability
- Results cover entire range of shareability:
  - users with local files only
  - users with global files only
- Little correlation among categories

Summary & Future Work

- Majority of surveyed, relevant files are shareable!
- Important component: private files as distinct from local files!
- Continuing survey, improving application
- Related project: Graffiti [Maltzahn07]