

# Searching Provenance

Shankar Pasupathy, Network Appliance

PASS Workshop, Harvard

October 2005

# Outline

- Why does Netapp care ?
- Questions we'd like to ask
- Use Google to search provenance ?
- Novel uses of searching provenance
- Distributed search of provenance
- Metrics

# Why does Netapp care ?

- Workflows (ILM)
  - Morgan Stanley wants to track how financial reports are generated
  - Set backup policies on workflows
- Generic search
  - Relevance of a document when ranking results
    - What sources were used to create the document ?
    - How authoritative are the sources ?
- Audit trails
- Don't backup stuff that can be easily recreated
  - E.g. object files (.o)
  - Thumbnails of images

# Questions we'd like to ask

- Forward and reverse queries
  - Given file “foo”, tell me the workflow that it's part of
  - What are all the data sets I can recreate easily from “foo” (need the notion of how long it takes to create descendants of foo)
    - How much space will I save ?
  - Given file “foo”, tell me exactly how to recreate it
- Fine-grain query
  - Which parts of other files (offset, length) were used to create “foo”

# Use Google ?

- Use Google to index all provenance information out there
  - No security (we could fix that)
  - Is it too heavy weight ?
- Are queries well structured ?
  - Perhaps SQL and a relational database work well

# Query mechanisms

- What the user sees
  - How would you modify NFS/CIFS to support provenance queries ?
  - Could you do something interesting via the filesystem
    - E.g. lookup an extended attribute whose name is well known to get the provenance tree
  - Modify fstat ?
- Visual representation of provenance
- Notification of provenance changes

# Query language...

- User access rights
  - Is uid/gid enough ?
  - Probably want to be specify roles and domains that the questioner has control over.
  - What if you don't have permissions to view a region of the provenance tree ?
- What about the provenance system itself
  - Is SQL good enough ?
  - RDF ontology/SPARQL query language

# Other uses of provenance

- Determine trust, authority of authors of documents
  - A provenance tree is similar to a paper citation tree
  - Who's cited most often ?
  - Use that to improve search results
- Craig Soules' work
  - Capture relationships among files over a period of time

# Distributed querying of provenance

- Data may have been produced from sources on different computers
- Distributed querying => common format for recording provenance
  - What is that format ?
- How do you describe partial or incomplete answers ?
  - Because parts of the distributed provenance tree are not available

# Metrics

- How do you compare PASS query systems ?
  - Performance
  - Relevance
  - Benchmarks