Missing Data in Large Transaction Databases

Allan R. Wilks
AT&T Labs - Research

Setting

- call detail on AT&T long distance network
- 300 million transactions (50 GB) per day
- collected from 400 sources
- reporting frequency ranging from continuous to every few weeks
- complicated variable-length record format

Use

- fraud detection
- streaming access
- database access

Problem

- are we seeing all the data?
- needle absence in haystack
- niches for fraudsters
- perception: database confidence

Sources

- are all sources reporting?
- depends on having exhaustive source list
- each source reporting everything?
  - volume monitoring
  - frequency monitoring
  - serial number monitoring
  - stratified -- all exchanges?

Holes in database

- users can detect quite small holes -- surprising
  - do users alert? -- depends on their expectations
  - do users think about the data as they see it?
- auto queries
  - transverse to reporting sources
- traceback
  - can the source of a hole be traced?
  - keep raw data

Tools

- streaming tools
  - sh, awk, C, ...
  - everything small
- database tools
  - Daytona
  - integrates well with UNIX
  - 8 TB and growing
- alerting
  - via pager
  - software failures
  - system failures
  - heartbeat

Lessons

- develop subject matter expertise
- log everything
- explain all anomalies
- keep raw data
- automate as much as possible