Collaboration Using OSSmole:
A repository of FLOSS data and analyses

Megan Conklin
Elon University

James Howison
Syracuse University

Kevin Crowston
Syracuse University
OSSmole Project

• Started as OSSmole in September, 2004
  
  But data was being collected as early as January, 2004 by the team

• We needed data about FLOSS projects

• 2 initial needs were:
  
  ★ Conklin: studying social networks and project team size as a power law
  
  ★ Howison & Crowston, et al: studying team dynamics, social networks

http://ossmole.sf.net
Getting Data

- Where to get FLOSS data?
- Scrape/spider Sourceforge
  - Sourceforge has a variety of reasons for not providing data dumps to the general research community (PII is big one)
  - Lots of individual spidering efforts going on

http://ossmole.sf.net
Sourceforge

- It's **big** — 100k
- It's relatively **stable**
- It's **slow** — `sleep(5)`

[Sourceforge URL](http://ossmole.sf.net)
Problems with Data

The problems with scraping Sourceforge are legion:

- **Practical** – SF bans IPs when it detects "bad" behavior
- **Inconsistencies** – SF can change HTML, data model
- **Economies of Scale** – Multiple efforts and no sharing
- **Lack of transparency** for methods used by a research team to gather their data, peer review is in question

http://ossmole.sf.net
OSSmole as a solution

The main goals of OSSmole are to be:

- **Collaborative** – data donations
- **Available** – encourage play!
- **Comprehensive** – should support ...
  - multiple repositories
  - historical data
- **Transparent** – data provenance...
  - model & data should be labeled & well-described (including origins)
- **High-Quality** – researchers can...
  - trust
  - verify
  - reproduce

http://ossmole.sf.net
Pretty Pictures

With OSSmole data, we've been able to do things like this:

the OSSmole project

http://ossmole.sf.net
More Pretty Pictures

...or this:

the OSSmole project  http://ossmole.sf.net
More Pretty Pictures

...and even this:

the OSSmole project  http://ossmole.sf.net
So the point is...

- At least we have the data now
- It's easy to find, and anyone can use it
But...

We're ready for v.2

- **Collaborative** -> encourage data donations
- **Available** -> live querying
- **Comprehensive** -> multiple repositories
  - freshmeat!!
  - savannah
  - stand-alone projects (important, large)
  - data from research papers?

- **New concern:** privacy of developers, change name (FLOSS?)

the OSSmole project  http://ossmole.sf.net
Next up...

☆ Freshmeat
  ✔ parse nightly RDF dump
  ☆ insert changes into db
  ☆ run some sample analyses (against SF?)

the OSSmole project

http://ossmole.sf.net