These slides were presented at the conference:

Oxford eResearch 2008

11-13 September 2008

University of Oxford

Conference website:
http://www.oii.ox.ac.uk/microsites/eresearch08/index.cfm

Conference papers collection:
http://ora.ouls.ox.ac.uk/objects/uuid%3A64aa6f39-7e81-4d42-a008-ee2d7524bd67

Conference organisers:

Oxford Internet Institute
University of Oxford
1 St Giles, Oxford OX1 3JS
http://www.oii.ox.ac.uk/
email: enquiries@oii.ox.ac.uk

Oxford e-Research Centre
University of Oxford
7 Keble Road, Oxford OX1 3QG
http://www.oerc.ox.ac.uk/

Replication of FLOSS Research as eResearch

Andrea Wiggins, James Howison, & Kevin Crowston
Syracuse University School of Information Studies

12 September 2008 ~ Oxford e-Research Conference

FLOSS Research

- Free/Libre Open Source Software
  - Software released under an “open source” license, often developed in a community-supported mode
- Research on FLOSS becoming more mainstream
  - Accessible example of other phenomena of interest
    - Virtual work
    - Self-organizing teams
    - Distributed collaboration
  - Data are readily available
    - FLOSS teams create and retain public archives of activities, providing great access for researchers

eResearch for FLOSS

- An approach to research using cyberinfrastructure, shared data and analyses to support collaboration
- FLOSS research is well suited to these approaches
  – Most researchers use the same raw data sources and have similar (or identical) data handling needs
  – Large volumes of shared research data are already available in "repositories of repositories", though this does not mean that they are easy to use
  – Social science application of research tools and practices that are currently more commonly used in natural and physical sciences

Replicating FLOSS Research

- Replicating a selection of FLOSS papers and presentations; efforts in progress
- Demonstrating utility and viability of eResearch approaches for FLOSS and social science
- Building reusable, customizable analysis components specific to FLOSS research, e.g. for data selection, sociomatrix generation for SNA, etc.
- Extending the original research analysis by implementing greater analysis flexibility and applying to new data sets

Studies Selected for Replication

Study	**Description**
Scozzi & Crowston, 2002	Applies competency rallying to predict success of projects based on various factors
Conklin, 2004	Examines distribution of project sizes for consistency with preferential attachment theory of growth in scale-free network
Howison et al., 2006	Examines dynamics of social networks of project communications over time
Robles et al., 2005	Examines growth rate of software
English & Schweik, 2007	Classifies projects based on metrics for success and stage of project growth

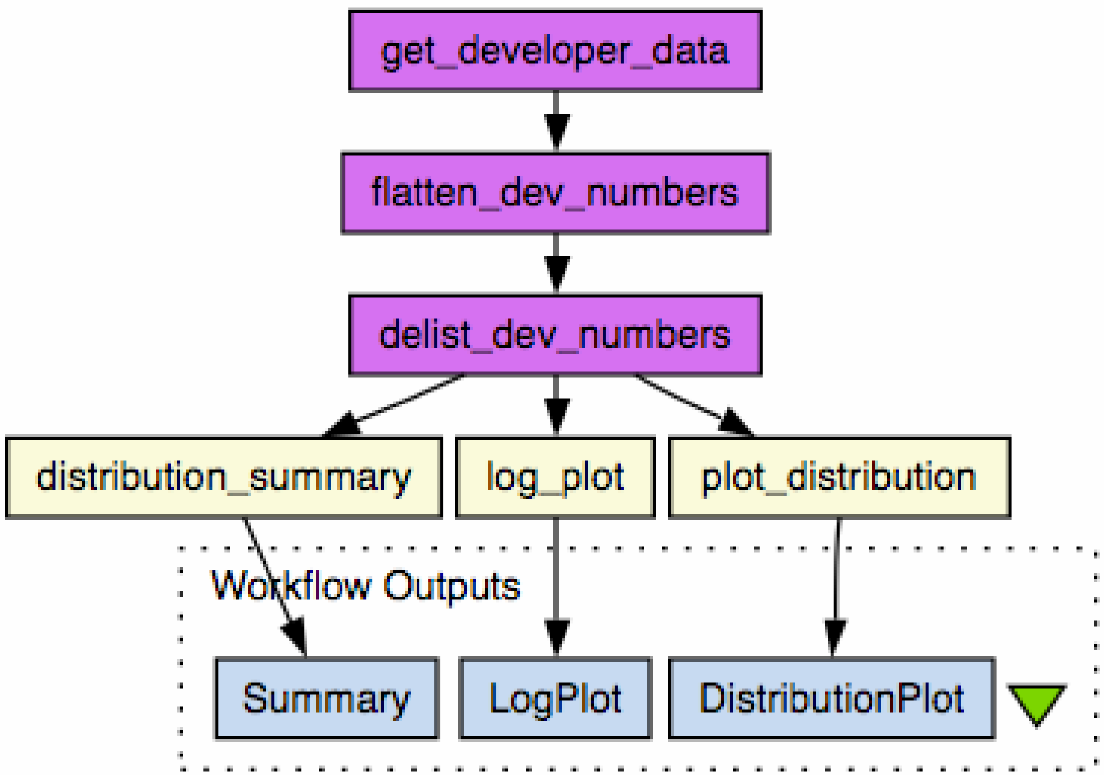
Using Taverna

- Scientific analysis workflow tool
  - Target users are UK life sciences community
- Create analysis workflows by connecting modular components through input/output ports
  - Produces analyses that are replicable, self-documenting, and easy to share
  - Components include WSDL SOAP web services, Beanshell, RShell, and local Java shims
- Collaboratively developing our workflows
  - Distributed efforts leveraging complementary skills

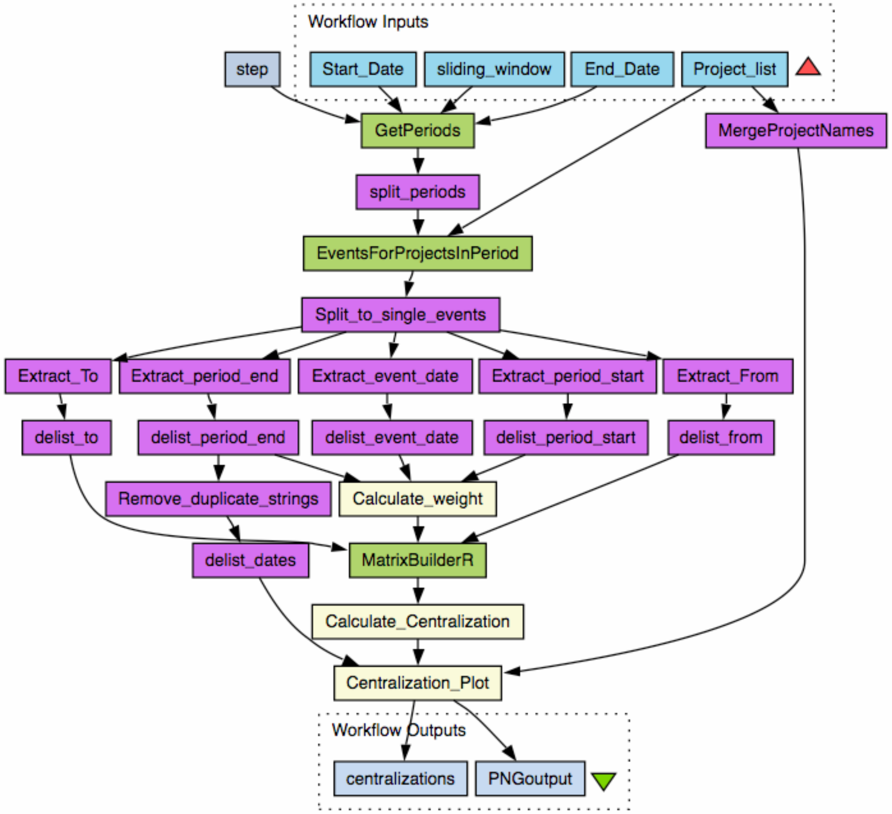
Building Workflows

1. Evaluate data, methods, analysis & findings in the selected papers
2. Specification of data requirements & list of desired outputs (may exceed original outputs)
3. Build abstract workflow representing expected analysis
4. Create/select components for precise operationalization of constructs, link them together
5. Iterative testing and development
6. Document, share, and use to explore analysis results

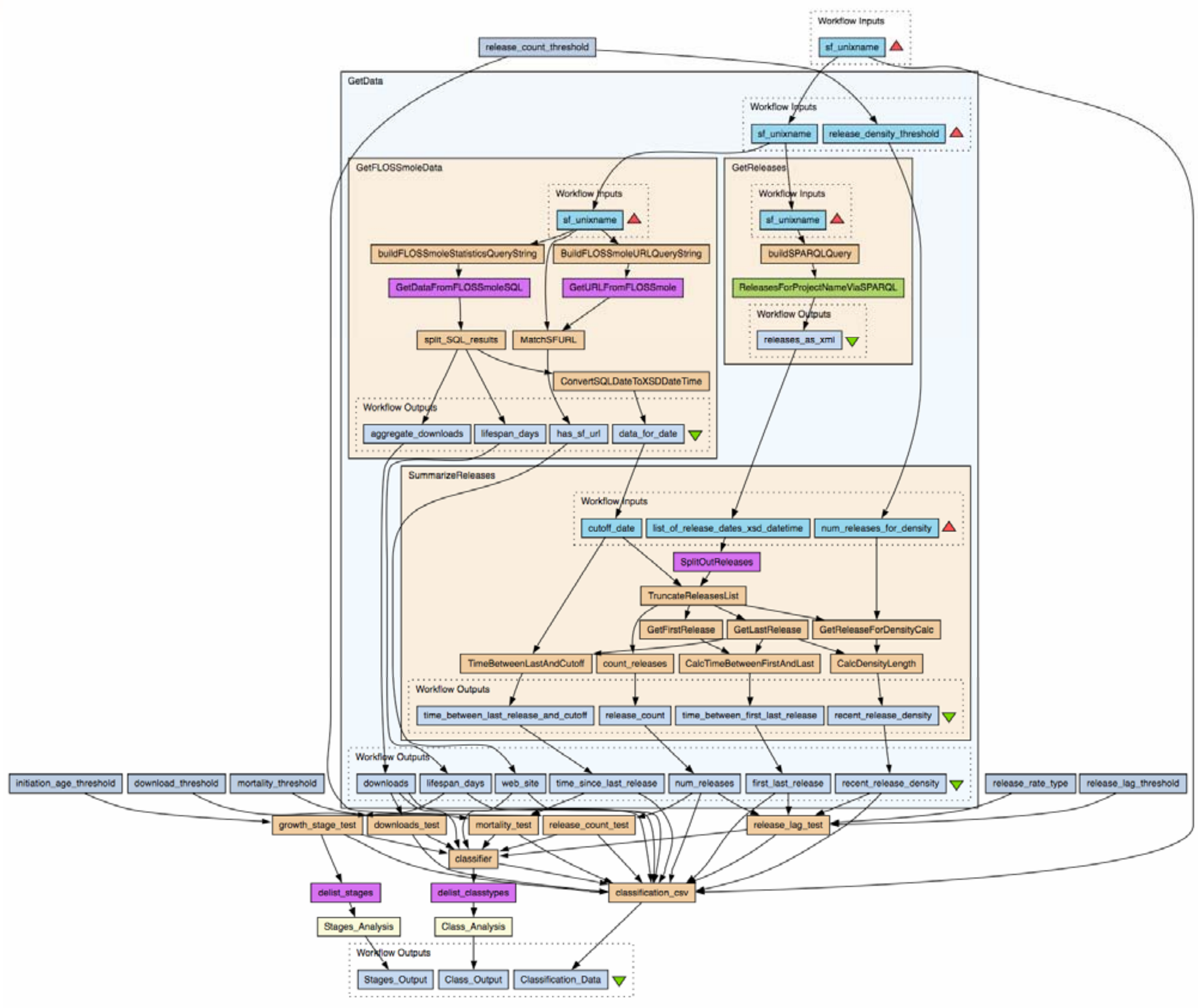
Do the Rich Get Richer?



Dynamic Social Network Analysis



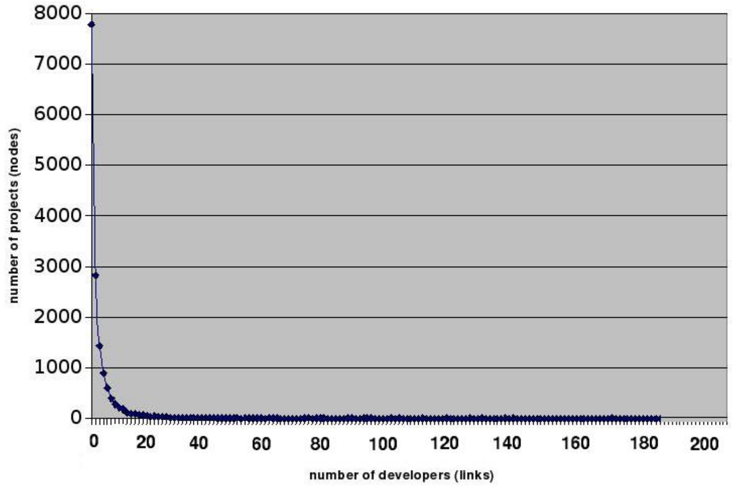
Classification of Projects

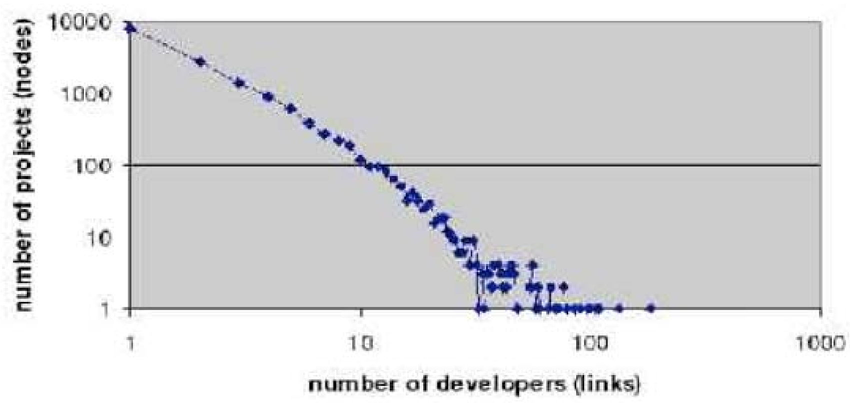


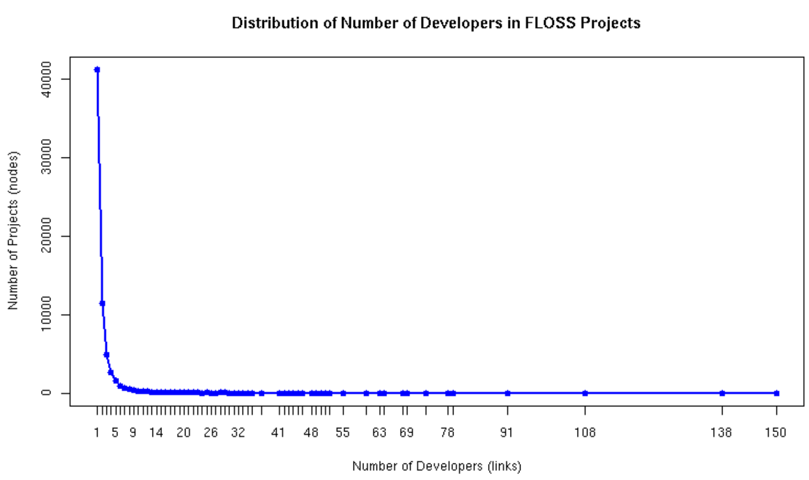
Using Workflows

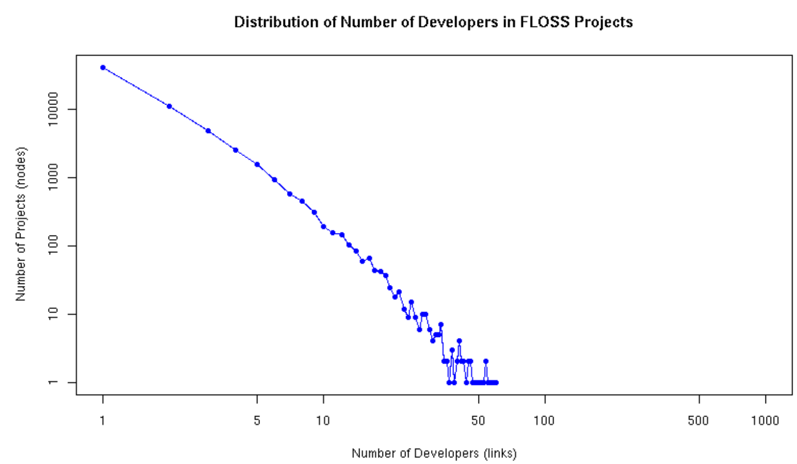
- Ran workflows on data samples appropriate for comparison to original work
- Extended the original work
  - Implemented future work suggestions from authors
  - Added in our own ideas
- Example: English & Schweick classification
  - Added two alternative approaches for a complex metric
  - Tested on known successful projects: classification shifted from success (original measure) to indeterminate (proposed measure)

Comparing Results

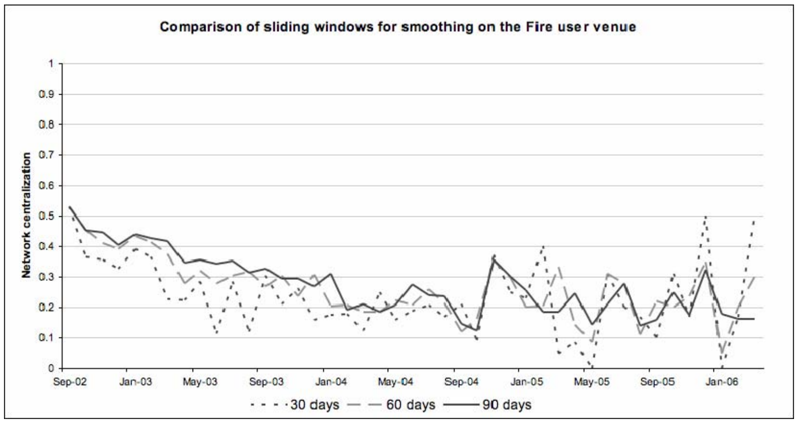


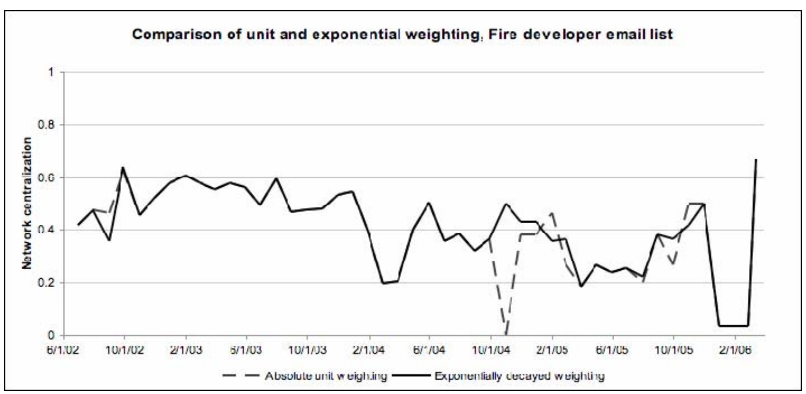






Comparing Analysis Parameters





- Examined effects of changing analysis parameters
  - Size of sliding window
  - Alternative edge weighting methods
- Expect to implement reusable workflow for exhaustive sensitivity testing
- Simplifies exploration of analysis parameters and rigorous validation of results

Sharing Workflows

Closing the loop between data and results by:

1. Documenting the workflow
   - Assign metadata to every describe component
   - Component code also contains comments
2. Eliminating potential web services security problems
   - Sample data set for public access
3. Sharing on MyExperiment.org with CC license
   - Link to/from related papers and data sets
   - http://www.myexperiment.org/groups/64

Lessons Learned: Data

- Replications required data from several RoRs
  - Modular design allows easy substitution of data retrieval components for future work with other sources
- One replication required data from multiple RoRs
  - Possible to mesh data, but not easy
  - Existence of consistent, persistent, unique project ID by forge is critical to successful data meshing
- Semantic technologies (e.g. RDF, OWL) show promise for improving interoperability across multiple RoRs without requiring structural changes to repository databases

Lessons Learned: Design

- Collaboratively developed analysis designs that emphasized flexibility and transparency
- Design benefits
  - Allows extension of research through analysis development and application to new data sources
  - Supports reuse of data and script resources
  - Preserves transparency for future reuse and auditing
- Goal: maximize potential uses of analysis workflow
  - Transparent structure and functionality
  - Reusable components
  - Low interdependence

Design Strategy: Parameterize

- Guiding design principle: parameterize all thresholds and variables
  - Enables sensitivity testing of a wide range of analysis factors
  - Incorporates more inputs while permitting maximum flexibility
- Allows straightforward implementation of alternative operationalization of concepts
  - Classification workflow example: rate of releases as an indicator of sustainable project activity
  - Implemented original measure plus two alternatives

Design Strategy: Modularity

- Guiding design principle: small things loosely connected
  - One operation or data manipulation per component
  - Components are logically complete, e.g. truth table
- Several notable benefits
  - Easier to co-develop components and integrate independent efforts
  - Can quickly change strategies with minimal adjustment of existing workflow structure
  - Can take advantage of a variety of component types
  - Easier to debug and execute modular components

Conclusions

- Shows the potential capacity for extending analysis and scale of research on FLOSS, demonstrates applicability of eResearch methods for the community
- Initial experiences are encouraging:
  - Workflow tools are flexible, and analyses can be tailored for reuse and extension
  - Despite the initial learning curve, using these methods greatly extended our analysis capacity
  - Greater adoption of these approaches requires addressing social factors, but has good potential for advancing the practice of social science research