

Analysis of the Open Source Software development community using ST mining: A Research Plan

Yongqin Gao, Greg Madey
Computer Science & Engineering
University of Notre Dame

NAACSOS Conference
Notre Dame, IN
June 26-28, 2005

Supported in part by the National Science
Foundation – ISS/Digital Science & Technology

Outline

- Background
- Motivation
- Problem definition
- Research data
- Methodology
- Conclusion

Background (OSS)

- What is OSS?
 - Free to use, modify and distribute
 - Source code available and modifiable
- Potential advantages over commercial software
 - Transparent and easy adoption
 - Fast development
 - Low cost
 - Potential high quality
- Why study OSS?
 - Software engineering — new development and coordination methods
 - Open content — model for other forms of open, shared collaboration
 - Complexity — successful example of self-organization/emergence
 - Growing popularity
 - Non-traditional governance and project management practices
 - Virtual --> Data!

SOURCEFORGE™
net


OpenOffice.org 1.0
The Open Source Office Suite

<http://www.freebsd.org>
FreeBSD
FreeBSD: The Power To Serve 

Open Source Software (OSS)

GNU



Savannah

- Free ...
 - to view source
 - to modify
 - to share
 - of cost

■ Examples

- Apache
- Perl
- GNU
- Linux
- Sendmail
- Python
- KDE
- GNOME
- Mozilla
- Thousands more



RePast



mozilla.org



The **Apache Software Foundation**

<http://www.apache.org/>

Python



Leaders



Linus Torvalds
Linux



Larry Wall
Perl



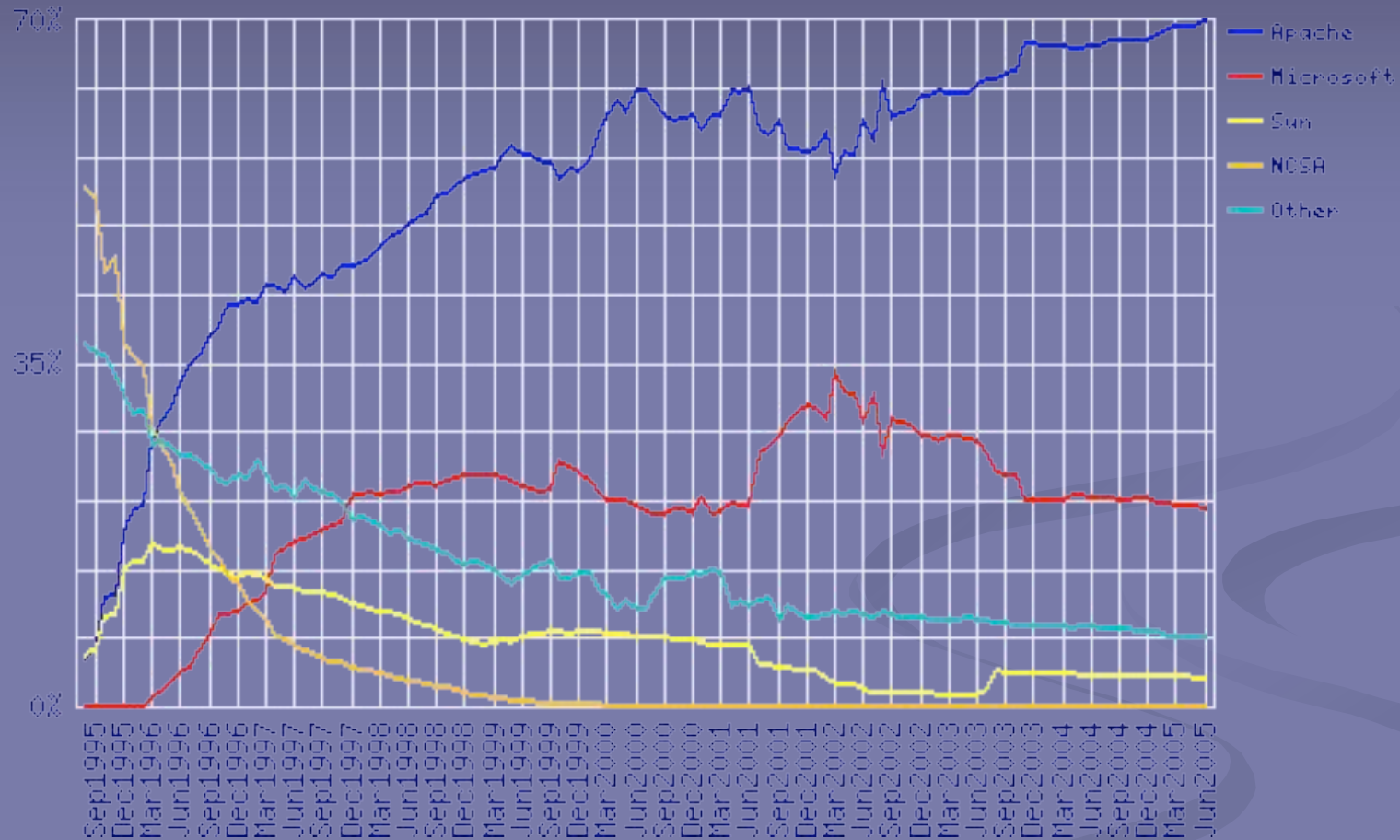
Eric Raymond
Cathedral and Bazaar



Richard M. Stallman

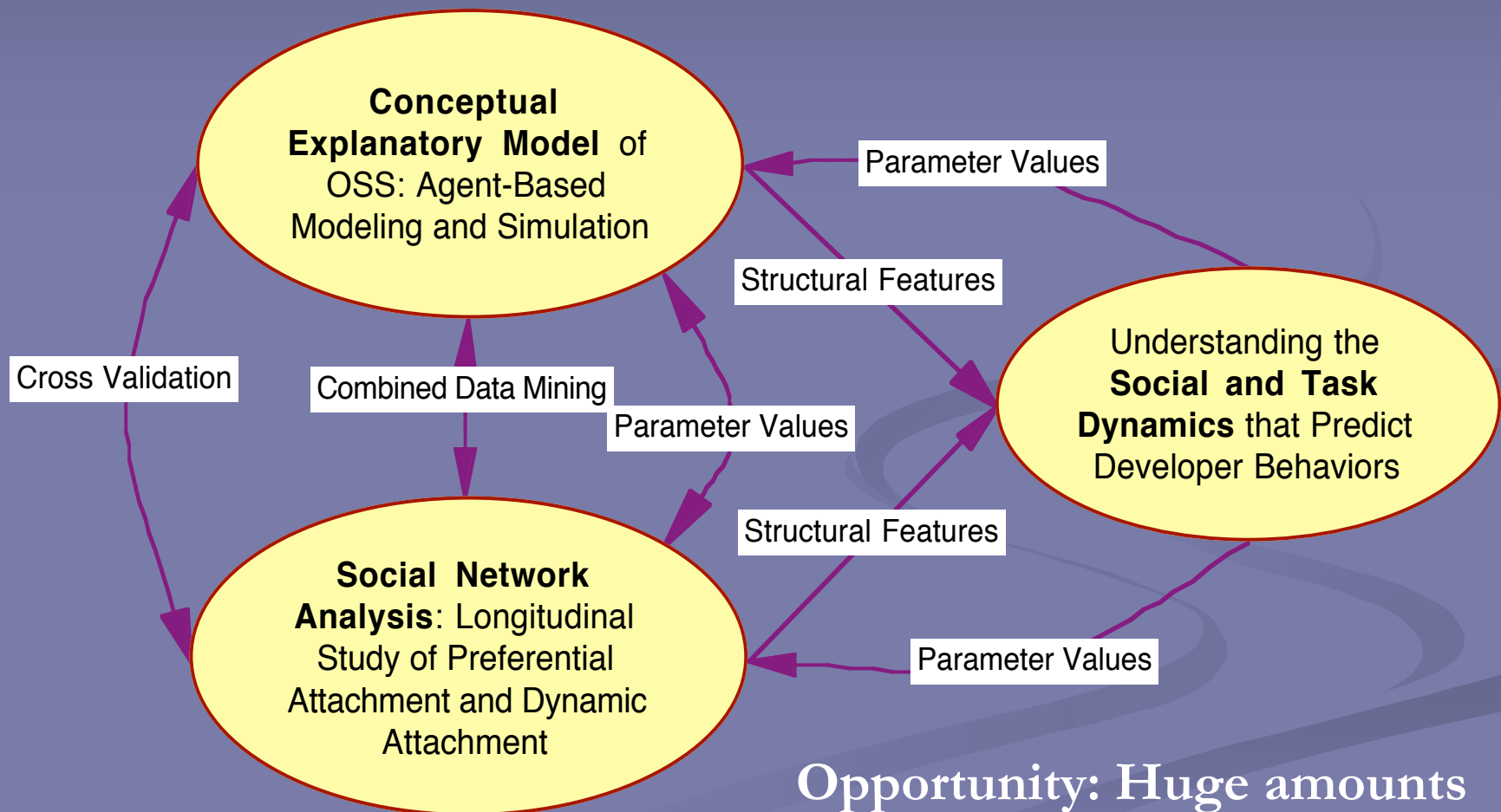
Richard Stallman
GNU Manifesto

Success of Apache



- Almost 70% Market Share (Netcraft.com)

Research Approach



Opportunity: Huge amounts of relatively good data

SourceForge.net

The screenshot shows a web browser window titled "@ SourceForge.net: Welcome". The address bar contains "http://sourceforge.net/". The browser's navigation bar includes buttons for Back, Forward, Stop, Refresh, Home, AutoFill, Print, and Mail. Below the address bar, there are several search engines and services listed, including Live Home Page, Apple, Apple Support, Apple Store, .Mac, Mac OS X, Microsoft, Office for Macintosh, and MSN. The main content area features a "Think Geek" banner at the top. Below the banner, there are navigation links for "my sf.net", "software map", "foundries", "about sf.net", and "My Favorites". The left sidebar contains a search box, "SF.net Resources" (Site Docs, Site Status, Site Map, Compile Farm, Project Help Wanted, New Releases, Contact SF.net Support), and "Most Active" projects (Compiere ERP + CRM Business Solution, Gaim, phpMyAdmin, XboxMediaPlayer). The main content area includes a "SourceForge.net is the world's largest Open Source software development website" section, a "Project of the Month" section (TUTOS), a "SourceForge.net Statistics" section (Hosted Projects: 58,685, Registered Users: 590,005), a "SourceForge.net Newsletter" sign-up form, and a "Latest News" section (Audacity sound editor 1.1.3 beta). A vertical banner on the right side of the page reads "www.devchannel.org".

- VA Software
- Part of OSDN
- Started 12/1999
- Collaboration tools
- 100 K Projects
- 100 K Developers
- 1 M Registered Users

150 GBytes of Data & Growing

SourceForge.net Research Data

Home

Overview

Papers

People

Research Data

SourceForge.net Research Data

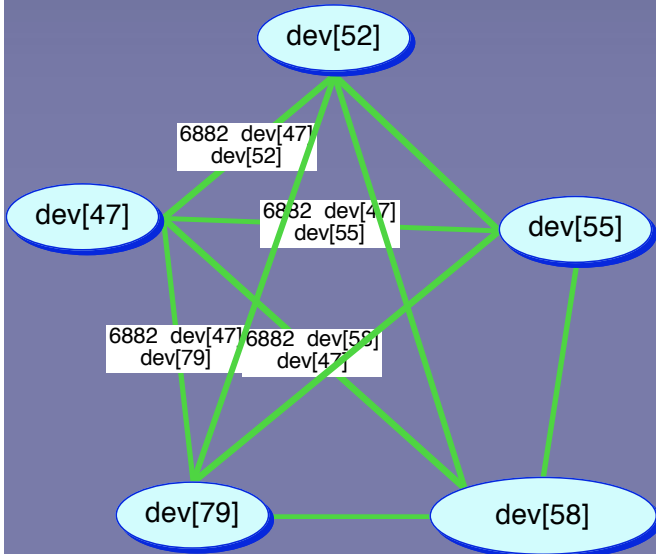
[SourceForge.net](#) is the world's largest Open Source software development web site, with the largest repository of Open Source code and applications available on the Internet. Owned and operated by OSTG, Inc. ("OSTG"), SourceForge.net provides free services to Open Source developers. The SourceForge.net web site is database driven and the supporting database includes historic and status statistics on approximately 100,000 projects and over 1 million registered users' activities at the project management web site. OSTG has shared certain SourceForge.net data with the University of Notre Dame for the sole purpose of supporting academic and scholarly research on the Free/Open Source Software phenomenon. OSTG has given Notre Dame permission to in turn share this data with other academic researchers studying the Free/Open Source Software phenomenon.

Release of the SourceForge.net Research Data

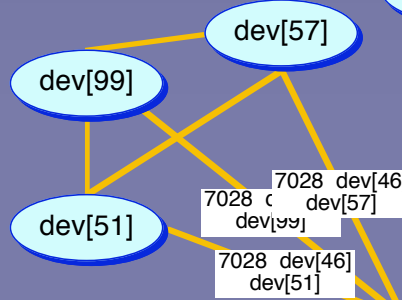
To advance the understanding of, and research on, the Free/Open Source Software phenomenon, portions of the data that may support such research, will be made available to academic or scholarly researchers. All requests for data must be submitted in writing ([e-mail](#)) to the Notre Dame PI, ([Greg Madey](#)). Only academic and scholarly researchers are eligible to receive the data. To receive the data, a short [questionnaire and agreement](#) must be completed, signed and returned.

OSS Developer - Social Network
 Developers are nodes / Projects are links
 24 Developers
 5 Projects
 2 Linchpin Developers
 1 Cluster

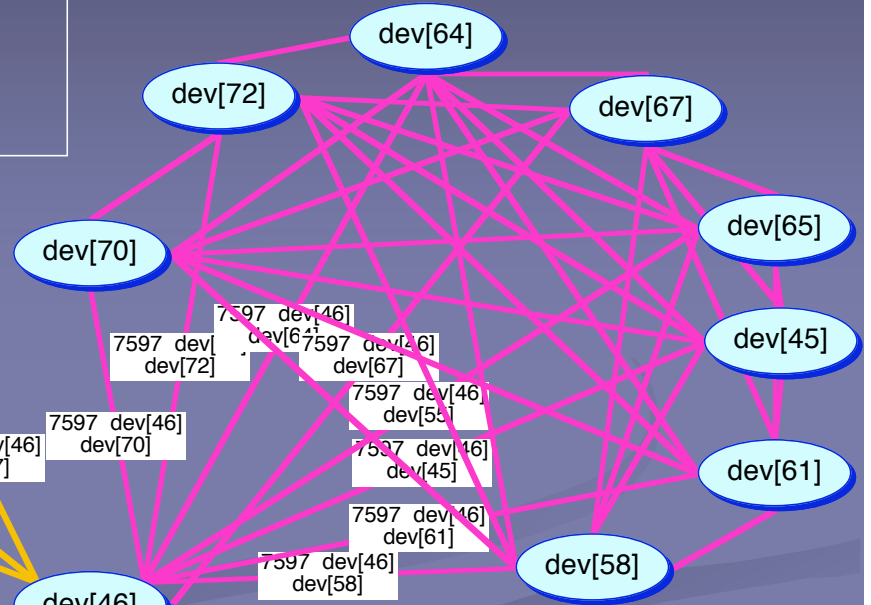
Project 6882



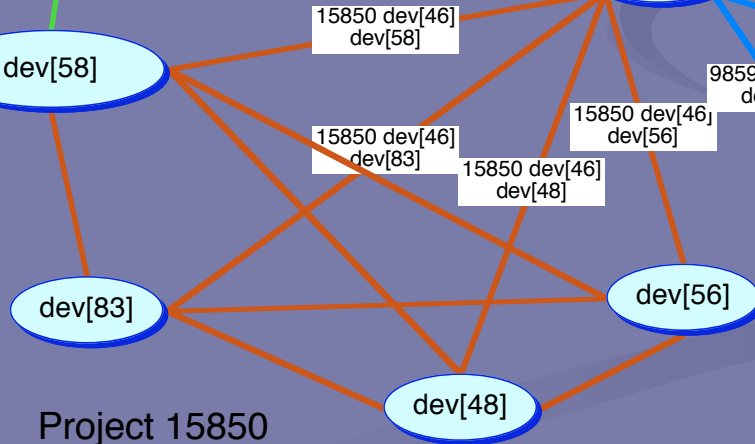
Project 7028



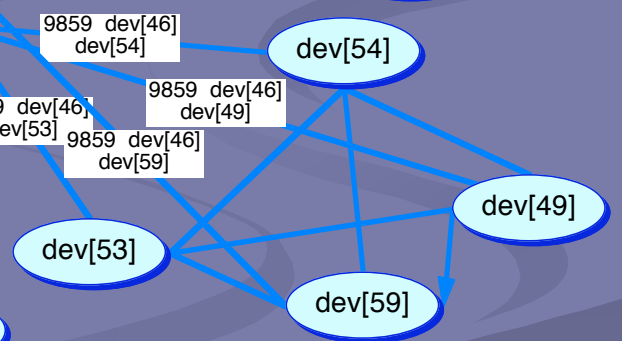
Project 7597



Project 15850



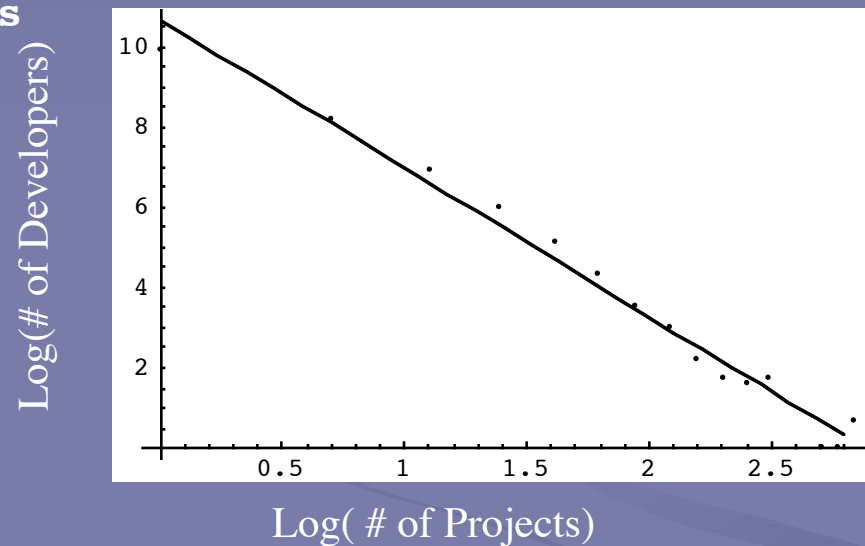
Project 9859



Scale free distribution: developer participation

projects # of developers on
that many projects

1	21488
2	3688
3	1086
4	413
5	177
6	76
7	35
8	21
9	9
10	6
11	5
12	6
15	1
16	1
17	1

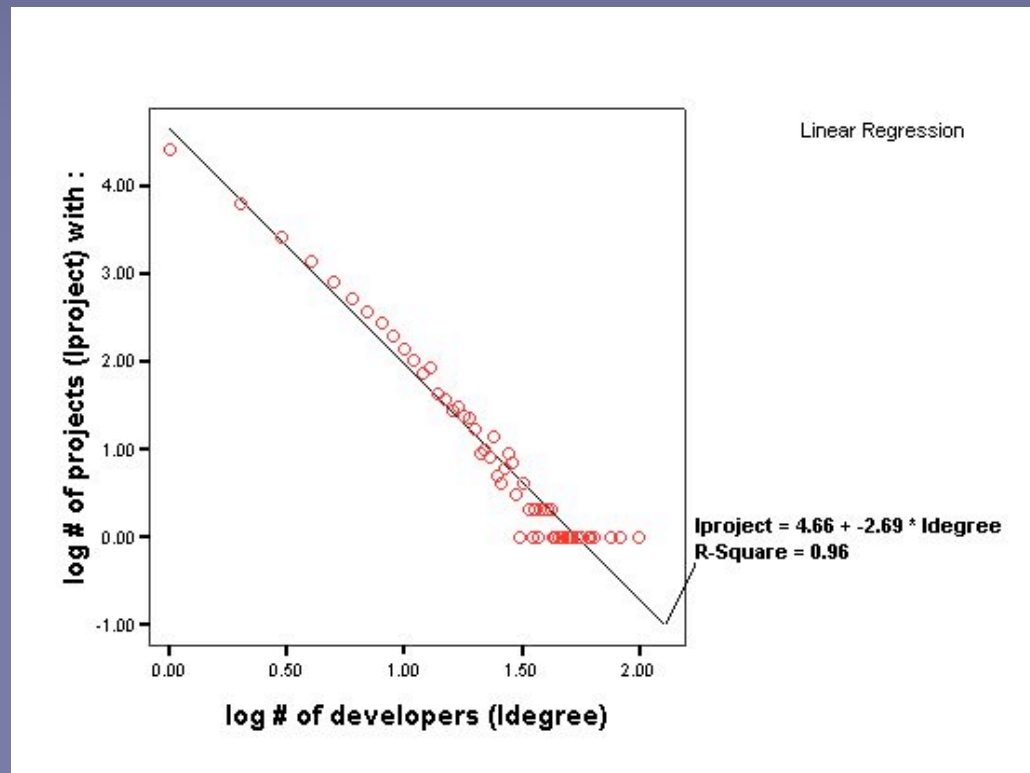


$$y = 10.6905 - 3.70892 x$$

$$R^2 = 0.979906$$

Scale Free – Power Law (developers)

Scale free distribution: project sizes

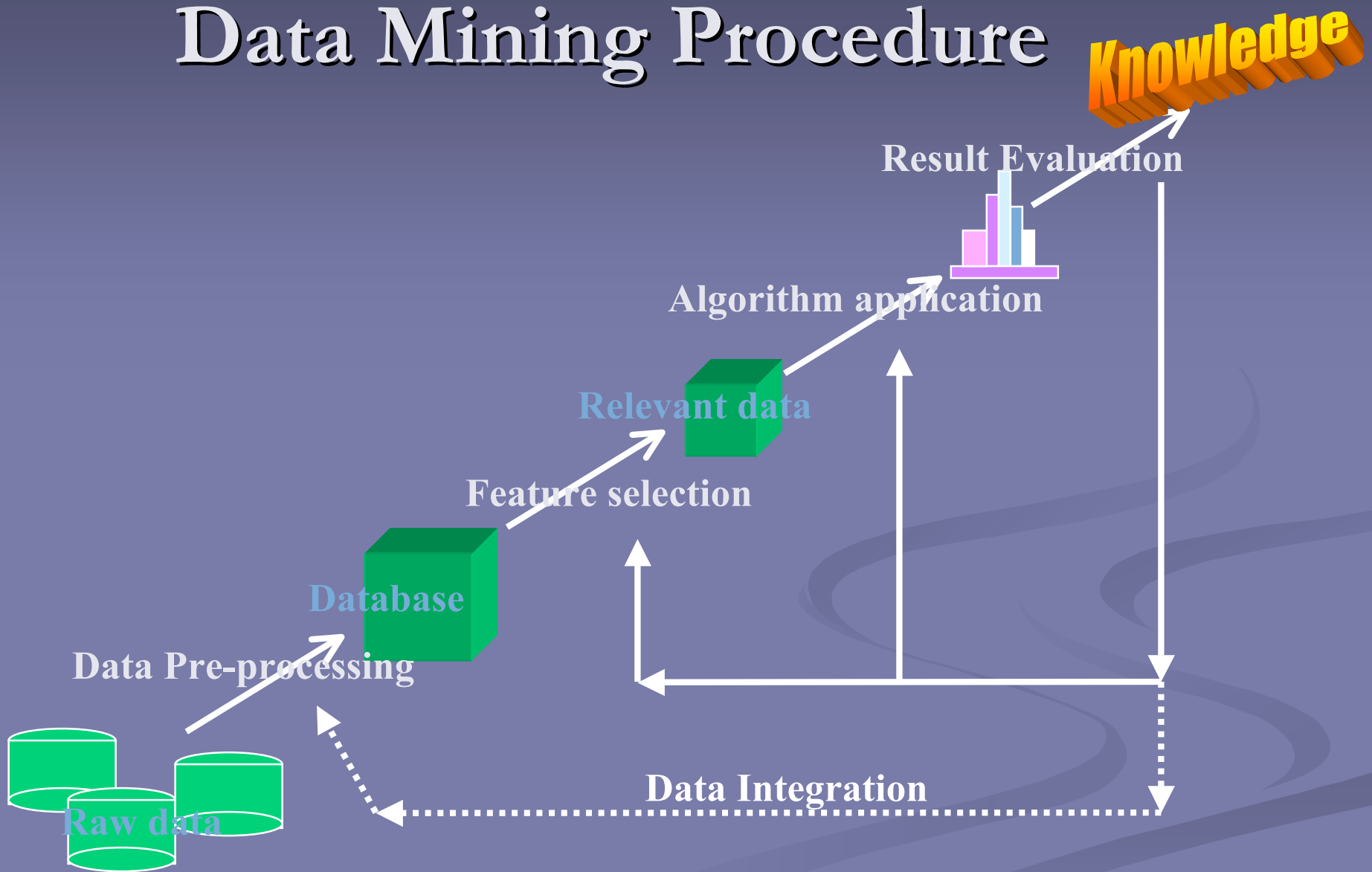


Scale Free – Power Law (projects)

Background (DM)

- Characteristics of data set
 - Incomplete, noisy, redundant
 - Complex structures, unstructured
 - Heterogeneous
 - Database not designed for research, but to support project management services of SourceForge.net
 - Temporal data is available, but not everything a researcher would want
 - Inferencing/discovery of temporal data potentially valuable opportunity
- What is DM (Data mining)
 - Nontrivial extraction of implicit, previously unknown and potentially useful information from data.

Data Mining Procedure



Spatial-temporal DM (1)

- Temporal data mining
 - Discover the behavior-based knowledge instead of state-based knowledge.
 - Example: many wolves -> fewer rabbits
 - Relationship between timely feedback and quality of software/success of the OSS project

Spatio-temporal DM

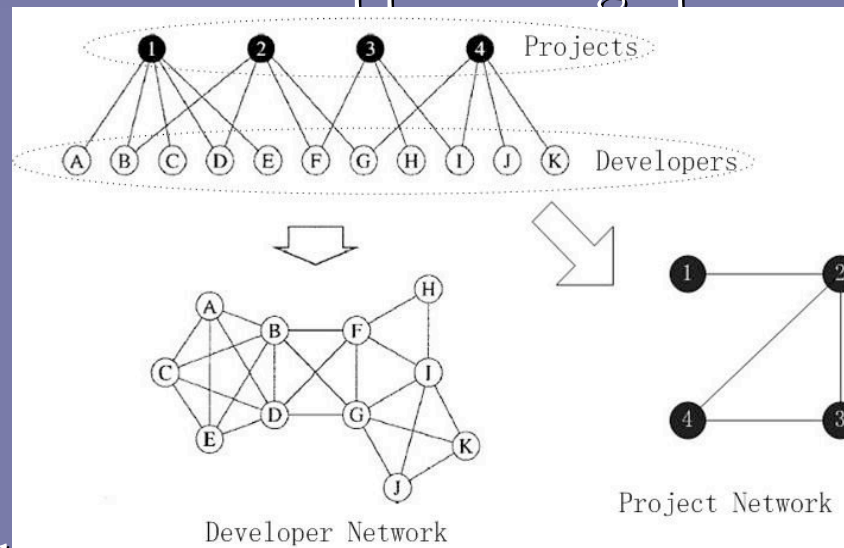
- New research domain: Spatio-temporal data mining
 - Growing interest in spatio-temporal data mining
 - Recommender systems
 - Location based services
 - Time based services
 - GIS applications
 - Extension of classic data mining techniques into data set with spatial and temporal properties.
 - Challenges: complexity of spatial information and difficulty in reasoning temporal information, e.g.,
 - Intervals
 - Points
 - Hybrids

Motivations

- Limitations of OSS research to date
 - Mostly feature based data mining to date
 - Neglecting of the inherent spatial and temporal information in the OSS community
- SourceForge.net properties
 - Spatial information
 - Collaboration network
 - Temporal information
 - History data and log tables

Spatial information in OSS?

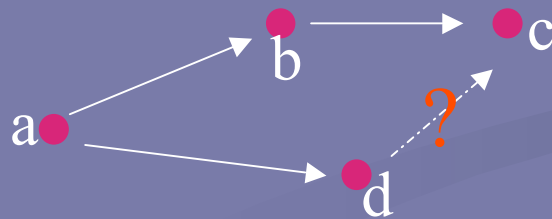
- The collaboration network in SF
 - Study of the topology of the collaboration network.
 - The network can be mapped as a graph



- This graph is a non-Metric space
- Spread of ideas (software engineering tools and practices, new project opportunities)

Temporal information in OSS

- The network is evolving and the histories of the site and individual entities comprise the temporal information in the network.
- Discrete time points
 - All the statistics are collected periodically.
- Partially ordered events
 - Multiple timelines existed in the system



ST Mining

- Different from classic data mining
 - Spatial and temporal relationships are complicated
 - Metric and non-metric spatial relations
 - Temporal relations
 - Intrinsic dependency and heterogeneity
 - Scale effect in space and time
- Significant modification of many data mining techniques are needed.

Problem definition I

- Dependency analysis
 - Extension of associations to ST mining
- Complicated associations
 - Vertical (temporal) and horizontal (spatial) associations
 - Combination of vertical and horizontal associations
 - Examples: lag effects between projects
- Flexible associations
 - Huge volume and scale effect of spatial-temporal data set introduce noise and error
 - Strict association is difficult to define

Problem definition II

- Topic of this study: prediction support
 - Clustering: group the projects with similar evolution.
 - Summarization: summarize the representative characteristics of different project evolution patterns
 - Prediction: predict the project evolution (based on the pattern discovered)

Research Data

- SourceForge.net database dump June 2005
 - 117 tables
 - Records up to 30 million per table
 - 23 Gigabytes
 - PostgreSQL
- Three types of tables
 - Data tables
 - History tables
 - Statistics tables

Methodology

- Project development statistics
 - Numerical statistics.
 - Expertise and survey statistics.
- Time series analysis
 - Generate the time series for these statistics
- Classification generation
 - ABN algorithm used
- Classifier evaluation
 - Evaluation by comparing the predicted class with the actual class

Numerical statistics

- Statistics tables have the information about project history
 - Stats_project_months
 - Every record stands for a monthly history of a single project
 - Records from November 1999 to June 2005
- There are 24 attributes in every record
 - Descriptive attributes (3)
 - Statistics (numeric) attributes (21)
- We use the statistics attributes

Statistics Attributes

Attributes	
Developers	Patches_opened
Downloads	Patches_closed
Subdomain_Views	Artifacts_opened
Page_views	Artifacts_closed
File_releases	Tasks_opened
Msg_posted	Tasks_closed
Bug_opened	Help_requests
Bug_closed	CVS_checkouts
Support_opened	CVS_commits
Site_views	CVS_adds
Support_closed	

Expertise statistics

- Rating scores
 - Expertise rating
 - User rating
- Importance parameter
 - Domain importance
 - Contribution parameter

Time Series

- Time series used to describe the history of each attribute.
 - Time series: an ordered sequence of values of a variable at equally spaced time intervals.
 - The available monthly values of each statistic is used to generate the time series.
- Goal is to study the project history patterns.
 - Description
 - Prediction

Conclusion

- Project prediction using ST mining
 - We used statistics to predict the project development
 - Calibration using new data is important to keep the prediction valid.

Questions