



**source**forge

**430K+ projects**

**3.7M+ users**

**GitHub**

**21.2M repositories**

**9M+ users**

 The Central Repository

**260K+ artifacts**

**70M+ downloads/week**

# Find, Create, and Publish Open Source software for free

THIS WEEK:

↓ 48,633,032 DOWNLOADS

✓ 376,453 CODE COMMITS

💬 3,367 FORUM POSTS

🐛 1,064 BUGS TRACKED

👁️ MORE DETAILS



FREE Quarterly Magazine



Featured Download

## IBM Bluemix - The Digital Innovation Platform

IBM Bluemix is a digital innovation platform designed to speed deployment & scale applications on the cloud



Learn More!



- Audio & Video
- Business & Enterprise
- Communications
- Development
- Home & Education
- Games
- Graphics
- Science & Engineering
- Security & Utilities
- System Administration

## Projects Of The Month



### Staff Choice Maxima -- GPL CAS based on DOE-MACSYMA

Computer Algebra System written in Common Lisp

Download

BSD | Windows | Linux



### Community Choice FlightGear - Flight Simulator

FlightGear Flight Simulator: free open-source multiplatform flight sim

Download


BSD | Windows | Mac | Linux

Editor's Choice

**MVNREPOSITORY**

Home » [org.springframework](#) » [spring-aop](#)

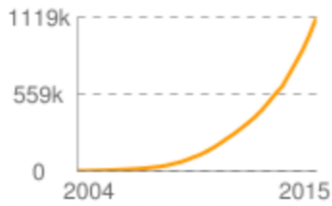
## Spring AOP

 **Spring AOP**  
[org.springframework](#) » [spring-aop](#) under **Aspect Oriented**

Spring AOP

Tags: [framework](#) [spring](#)

**Artifacts/Year**



**Popular Categories**

- Aspect Oriented
- Actor Frameworks

**The Central Repository** SEARCH | [ADVANCED SEARCH](#) | [BROWSE](#) | [QUICK STATS](#)

[New: About Central](#) | [Advanced Search](#) | [API Guide](#) | [Help](#)

### Browse Central For [org.openrdf.sesame : sesame : 4.0.0](#)

Click on a link above to browse the repository.

**Project Information**

GroupId:

ArtifactId:

Version:

**Dependency Information**

**Apache Maven**

```
<dependency>
  <groupId>org.openrdf.sesame</groupId>
  <artifactId>sesame</artifactId>
  <version>4.0.0</version>
</dependency>
```

[Apache Buildr](#)

[Apache Ivy](#)

[Groovy Grape](#)

[Gradle/Grails](#)

[Scala SBT](#)

[Leiningen](#)

**Project Object Model (POM)**

```
<project xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://www.w3.org/2001/X
  <modelVersion>4.0.0</modelVersion>
  <parent>
    <groupId>org.sonatype.oss</groupId>
    <artifactId>oss-parent</artifactId>
    <version>7</version>
  </parent>
  <groupId>org.openrdf.sesame</groupId>
  <artifactId>sesame</artifactId>
  <version>4.0.0</version>
  <packaging>pom</packaging>
  <name>OpenRDF Sesame</name>
  <description>An extensible framework for RDF and RDF Schema data.</description>
  <url>http://www.openrdf.org/</url>
  <inceptionYear>2001</inceptionYear>
  <organization>
    <name>Aduna</name>
    <url>http://www.aduna-software.com/</url>
  </organization>
  <developers>
```

## What's the code?

### Automatic Classification of Source Code Archives

Secil Ugurel<sup>1</sup>, Robert Krovetz<sup>2</sup>, C. Lee Giles<sup>1,2,3</sup>, David M. Pennock<sup>2</sup>,  
Eric Glover<sup>2</sup>, Hongyuan Zha<sup>1</sup>

### Automatic Categorization Algorithm for Evolvable Software Archive

Shinji Kawaguchi<sup>†</sup>, Pankaj K. Garg<sup>††</sup>  
Makoto Matsushita<sup>†</sup> and Katsuro Inoue<sup>†</sup>

## On using machine learning to automatically classify software applications into domain categories

Mario Linares-Vásquez • Collin McMillan •  
Denys Poshyvanyk • Mark Grechanik

### MUDABlue: An Automatic Categorization System for Open Source Repositories

Shinji Kawaguchi<sup>†</sup>, Pankaj K. Garg<sup>††</sup>, Makoto Matsushita<sup>†</sup> and Katsuro Inoue<sup>†</sup>  
†Graduate School of Information Science and Technology, Osaka University  
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan  
{s-kawagt, matusita, inoue}@ist.osaka-u.ac.jp  
††Zee Source  
1684 Nightingale Avenue, Suite 201  
Sunnyvale, California, 94807, USA  
garg@zeesource.net

### Categorizing Software Applications for Maintenance

Collin McMillan<sup>1</sup>, Mario Linares-Vásquez<sup>2</sup>, Denys Poshyvanyk<sup>1</sup>, Mark Grechanik<sup>3</sup>  
<sup>1</sup>Department of Computer Science  
The College of William and Mary  
Williamsburg, Virginia, USA  
{cmc, denys}@cs.wm.edu  
<sup>2</sup>Department of Computer Science  
Universidad Nacional de Colombia  
Bogotá, Colombia  
mlinaresv@unal.edu.co  
<sup>3</sup>Accenture Technology Labs and  
The University of Illinois at Chicago  
Chicago, Illinois, USA  
drmark@uic.edu

### Using Latent Dirichlet Allocation for Automatic Categorization of Software

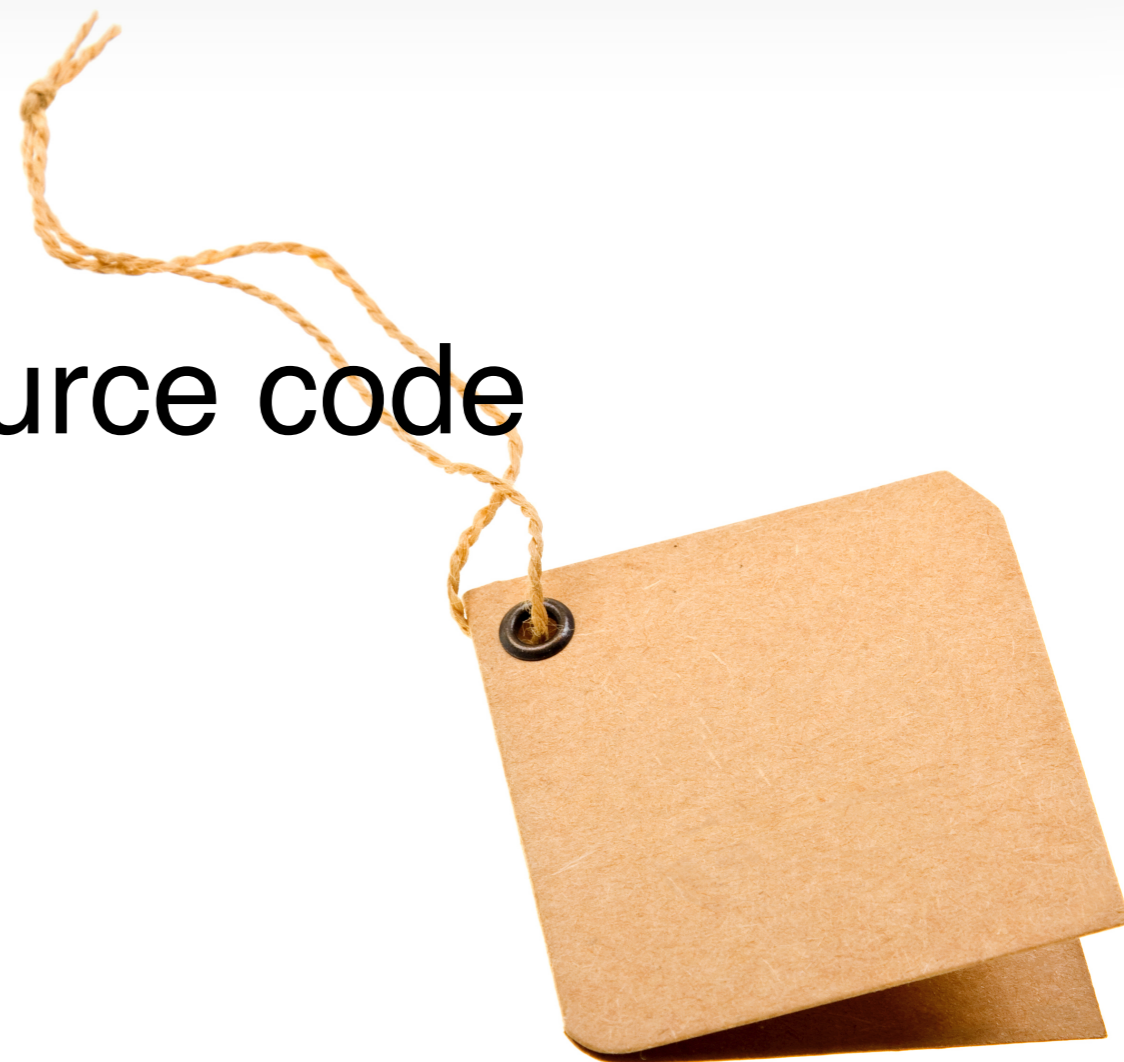
Kai Tian, Meghan Revelle, and Denys Poshyvanyk  
*Computer Science Department  
The College of William and Mary  
Williamsburg, VA 23185  
{ktian, meghan, denys}@cs.wm.edu*

# Current Issues

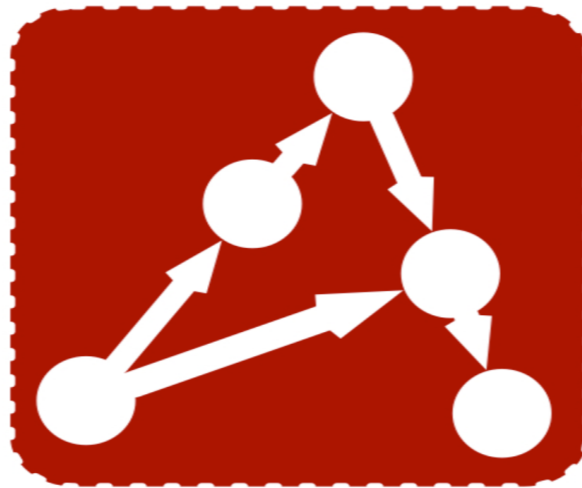
1. Source code is not always available
2. Predefined categories may not be sufficient
3. Supervised categorization is not always possible

# Automatic Categorization: Desired Features

1. Multi-label
2. Does not depend on source code
3. Unsupervised
4. Meaningful labels



**S**

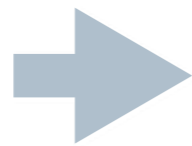


**LLY**

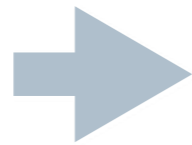
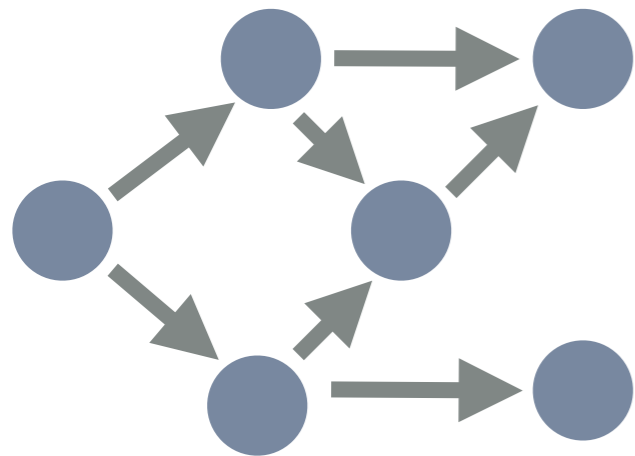


# SALLY

access  
accuracy  
acronym  
web  
weight  
white  
whitespace  
width  
wifi



Identifiers  
=  
domain



Dependencies  
=  
domain



PRIMARY  
AND  
SECONDARY  
TAGS  
+  
DEFINITIONS



[Home](#)

[Projects](#)

[Categories](#)

GroupId

ArtifactId

Filter

GroupId	ArtifactId	Version
<a href="#">org.openrdf.sesame</a>	<a href="#">sesame-model</a>	2.7.12
<a href="#">org.openrdf.sesame</a>	<a href="#">sesame-rio-api</a>	2.7.12
<a href="#">org.openrdf.sesame</a>	<a href="#">sesame-rio-binary</a>	2.7.12
<a href="#">org.openrdf.sesame</a>	<a href="#">sesame-rio-datatypes</a>	2.7.12
<a href="#">org.openrdf.sesame</a>	<a href="#">sesame-rio-languages</a>	2.7.12
<a href="#">org.openrdf.sesame</a>	<a href="#">sesame-rio-n3</a>	2.7.12
<a href="#">org.openrdf.sesame</a>	<a href="#">sesame-rio-nquads</a>	2.7.12
<a href="#">org.openrdf.sesame</a>	<a href="#">sesame-rio-ntriples</a>	2.7.12
<a href="#">org.openrdf.sesame</a>	<a href="#">sesame-rio-rdfjson</a>	2.7.12



[Home](#) [Projects](#) [Categories](#)

# of Tags

[org.openrdf.sesame.sesame-rio-rdfjson-2.7.12](#)



**rdf** 38% **json** 27% graph 7% literal 6% statement 6% model 4% report 4% datatype

3% fatal 2% config 2%

**rdf** 26% **datatype** 15% **owl** 10% statement 9% fail 7% rdfa 7% literal 7% language 6%

axiom 6% registry 6%

rdf



StackOverflow

Wikipedia

Techtarget

The Resource Description Framework (RDF) is a language for representing information about resources in the World Wide Web. It is a syntax independent data model that may be serialised in a variety of concrete syntaxes. RDF is the core data format used on the Semantic Web.

*Definition from StackOverflow.*

Close

**rdf** 38% **json** 27% graph 7% literal 6% statement 6% model 4% report 4% datatype

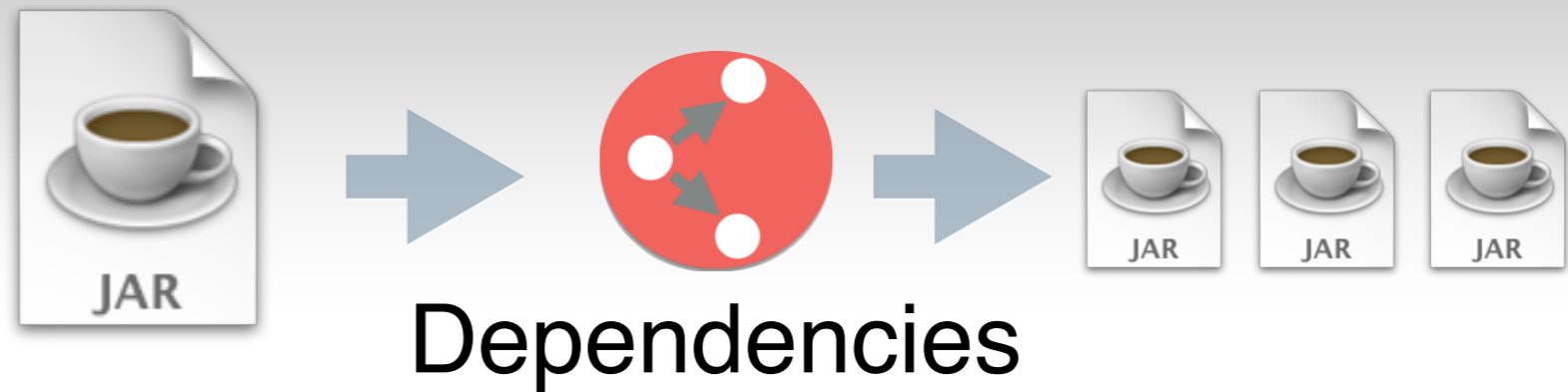
3% fatal 2% config 2%

**rdf** 26% **datatype** 15% **owl** 10% statement 9% fail 7% **rdfa** 7% literal 7% language 6%

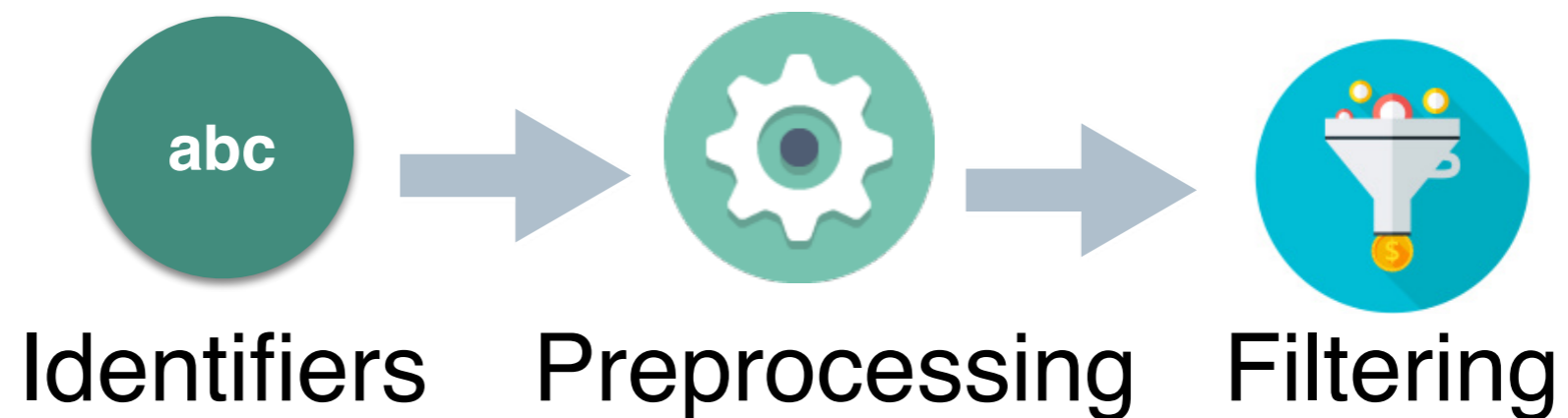
axiom 6% registry 6%

# SALLY

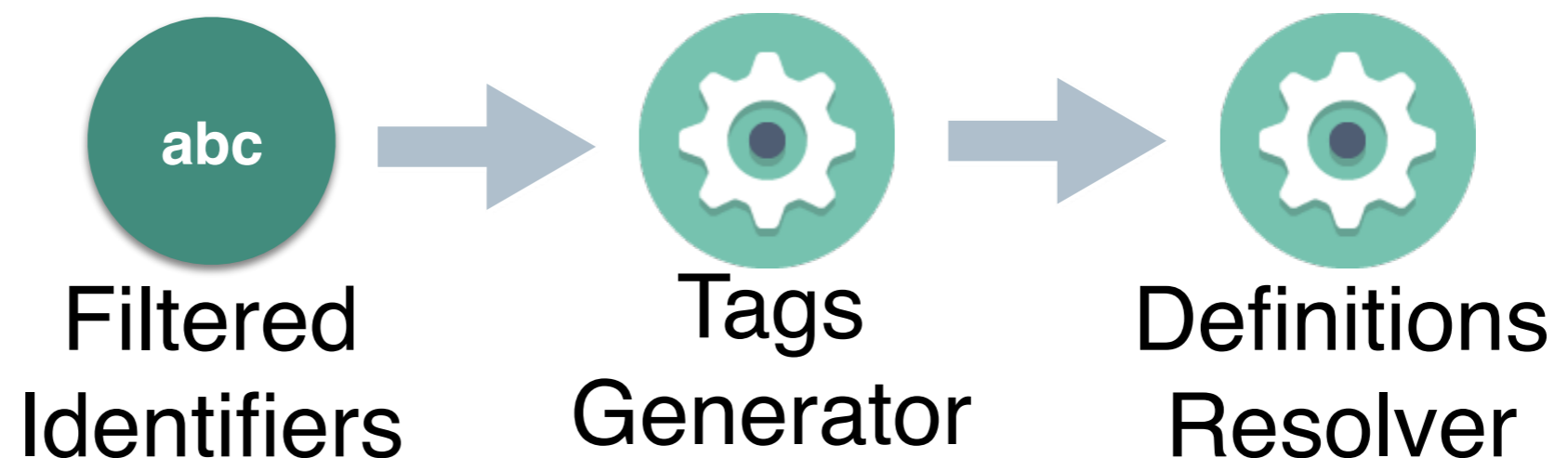
STEP 1



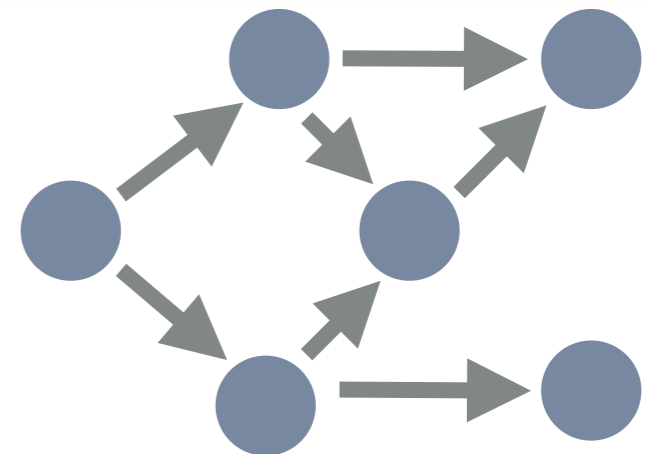
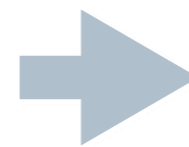
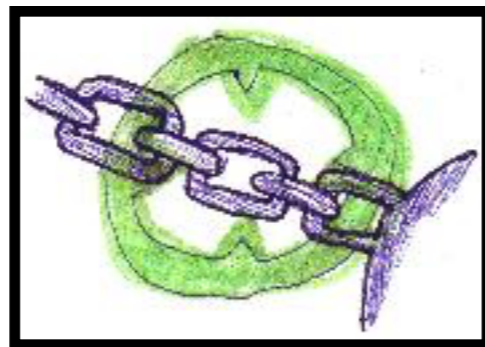
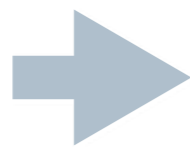
STEP 2



STEP 3

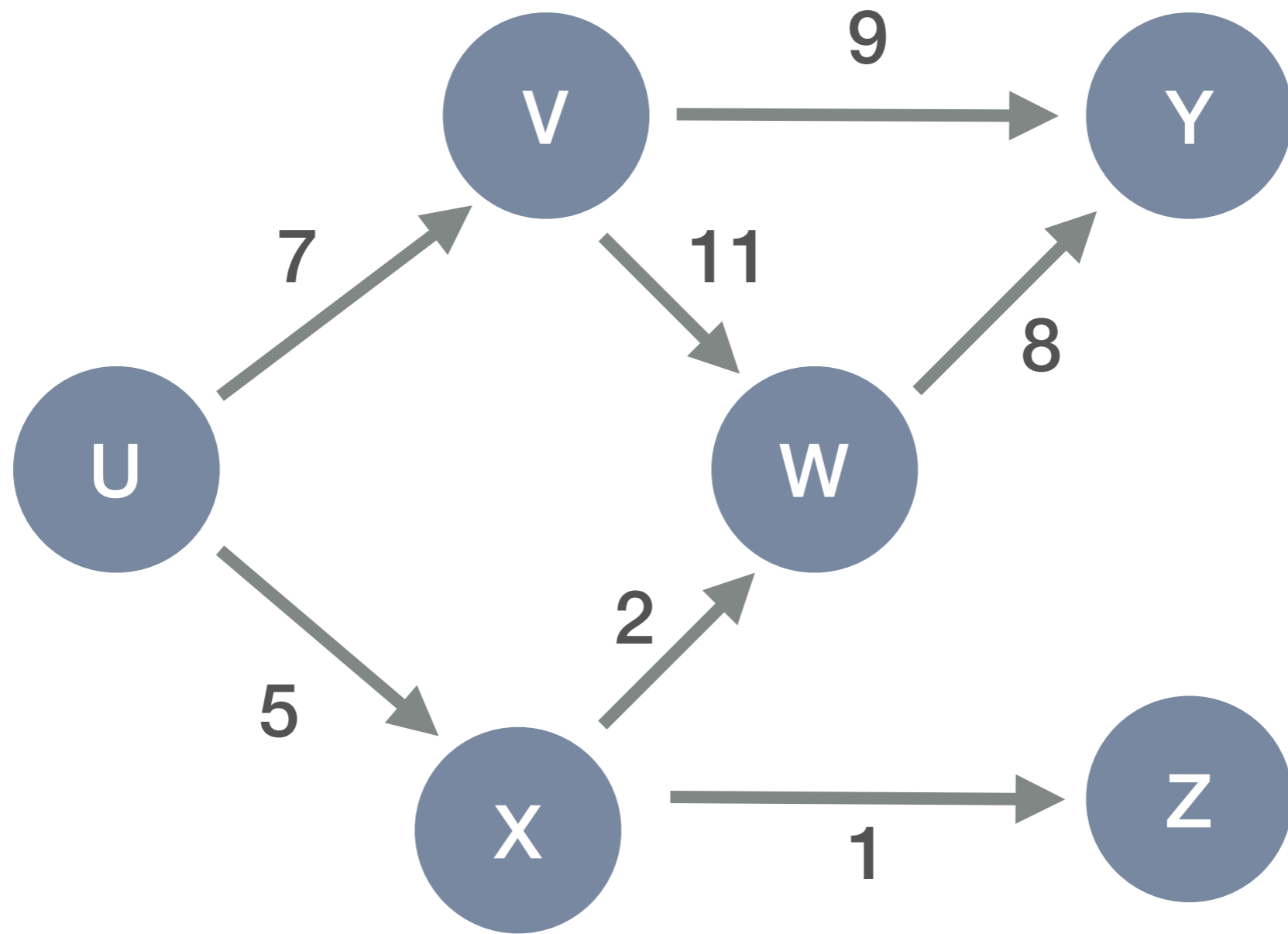


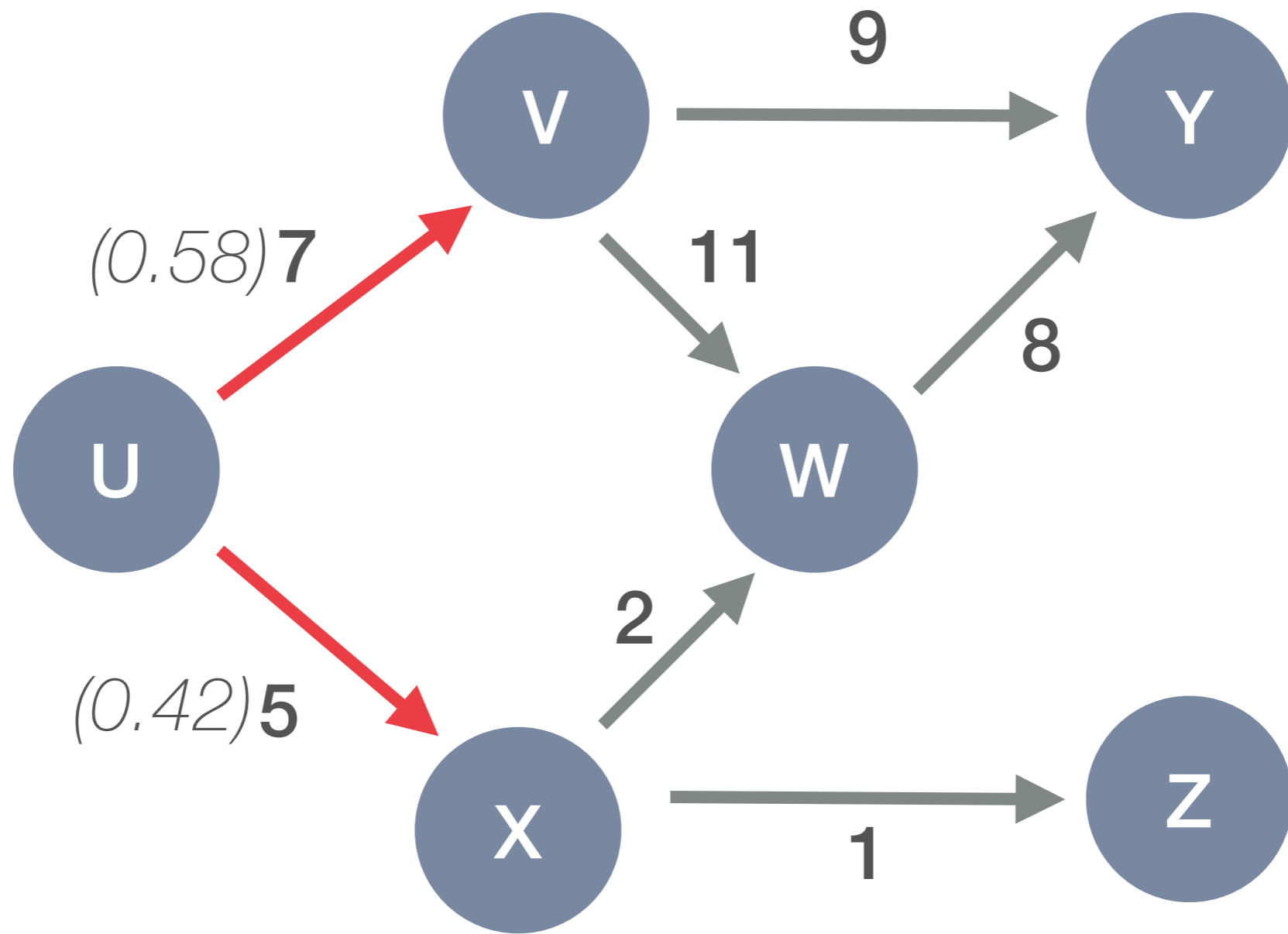
# SALLY: Dependencies



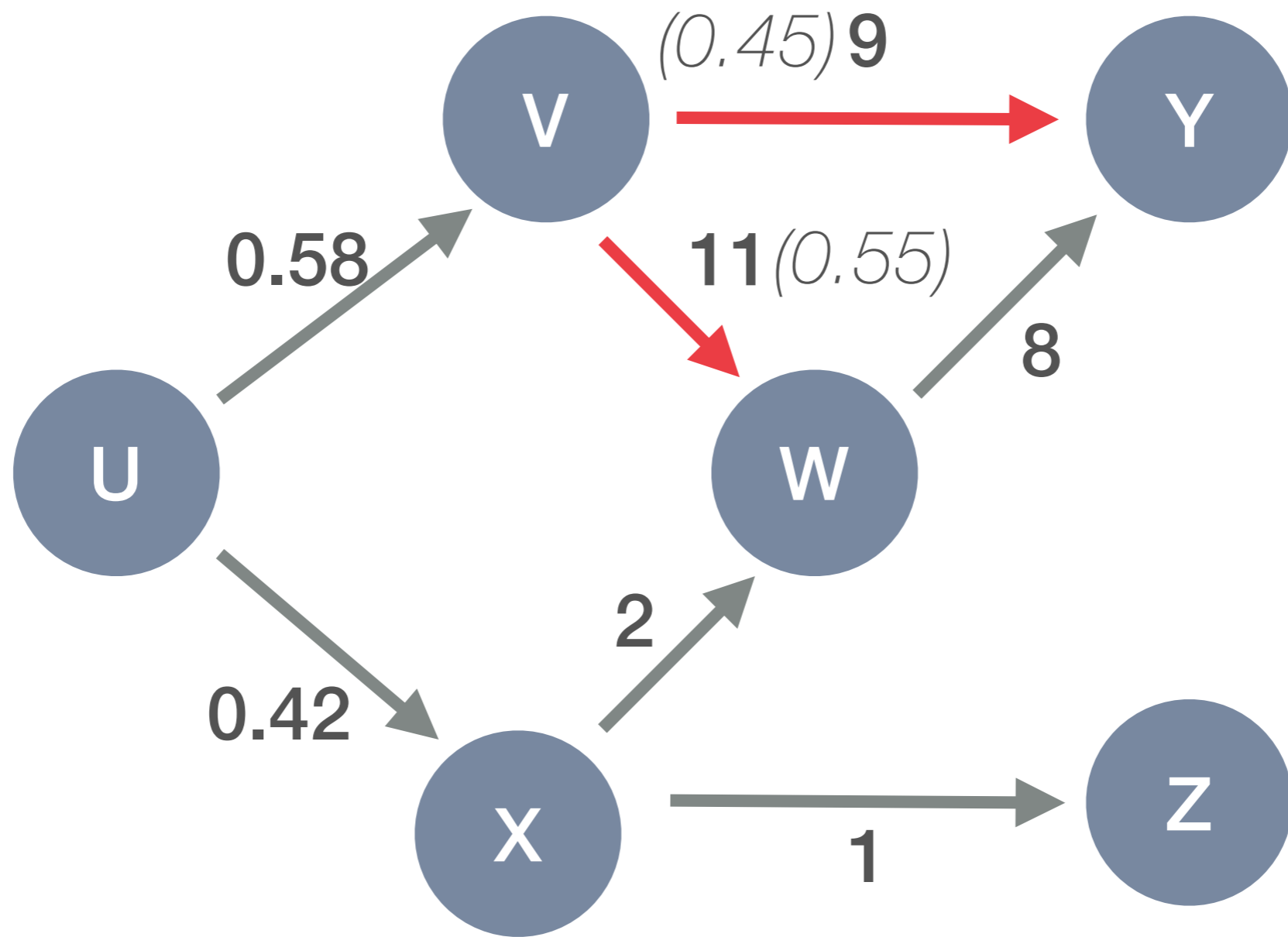
Dependency Finder\*

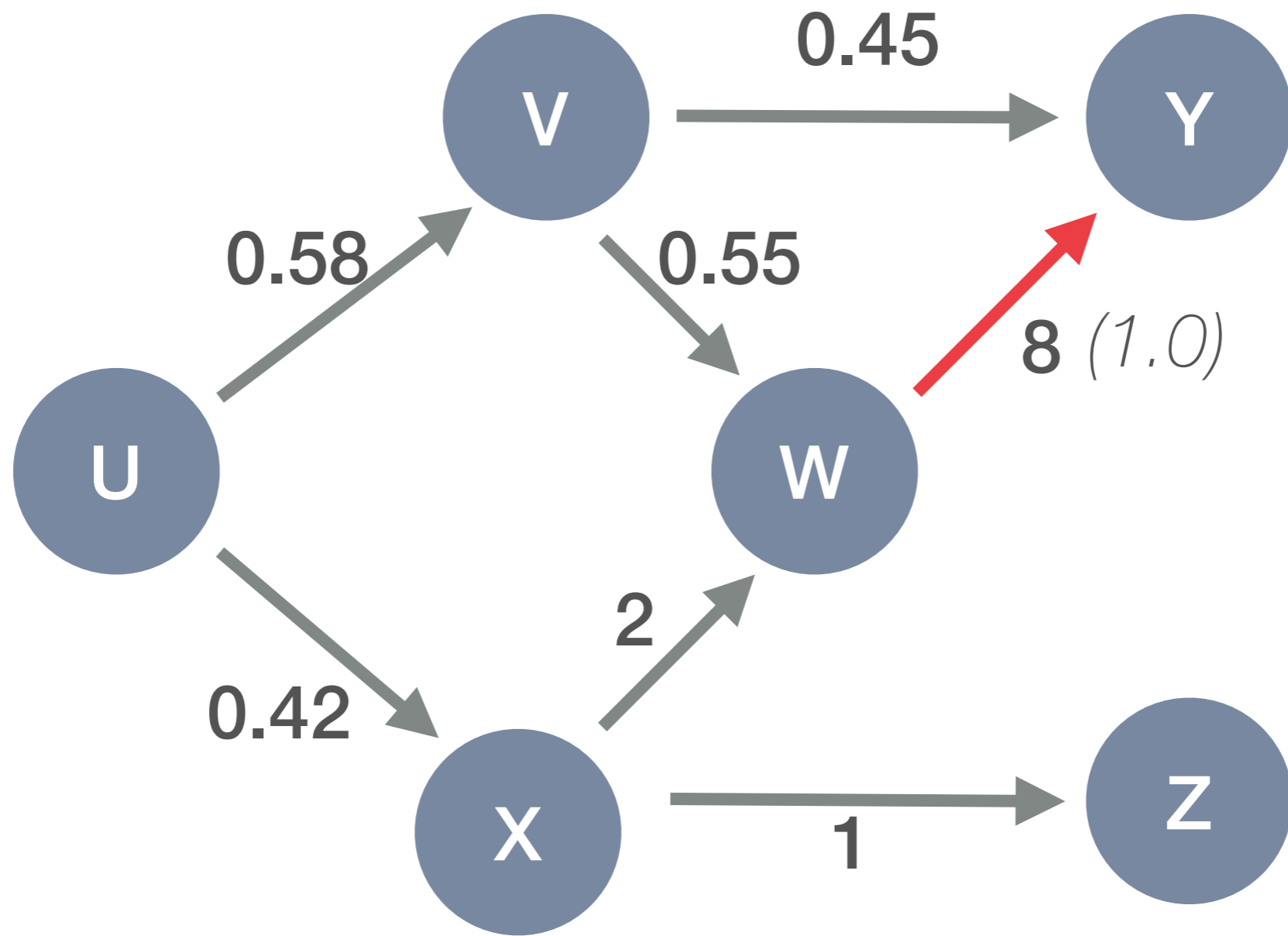
\*<http://deppfind.sourceforge.net>

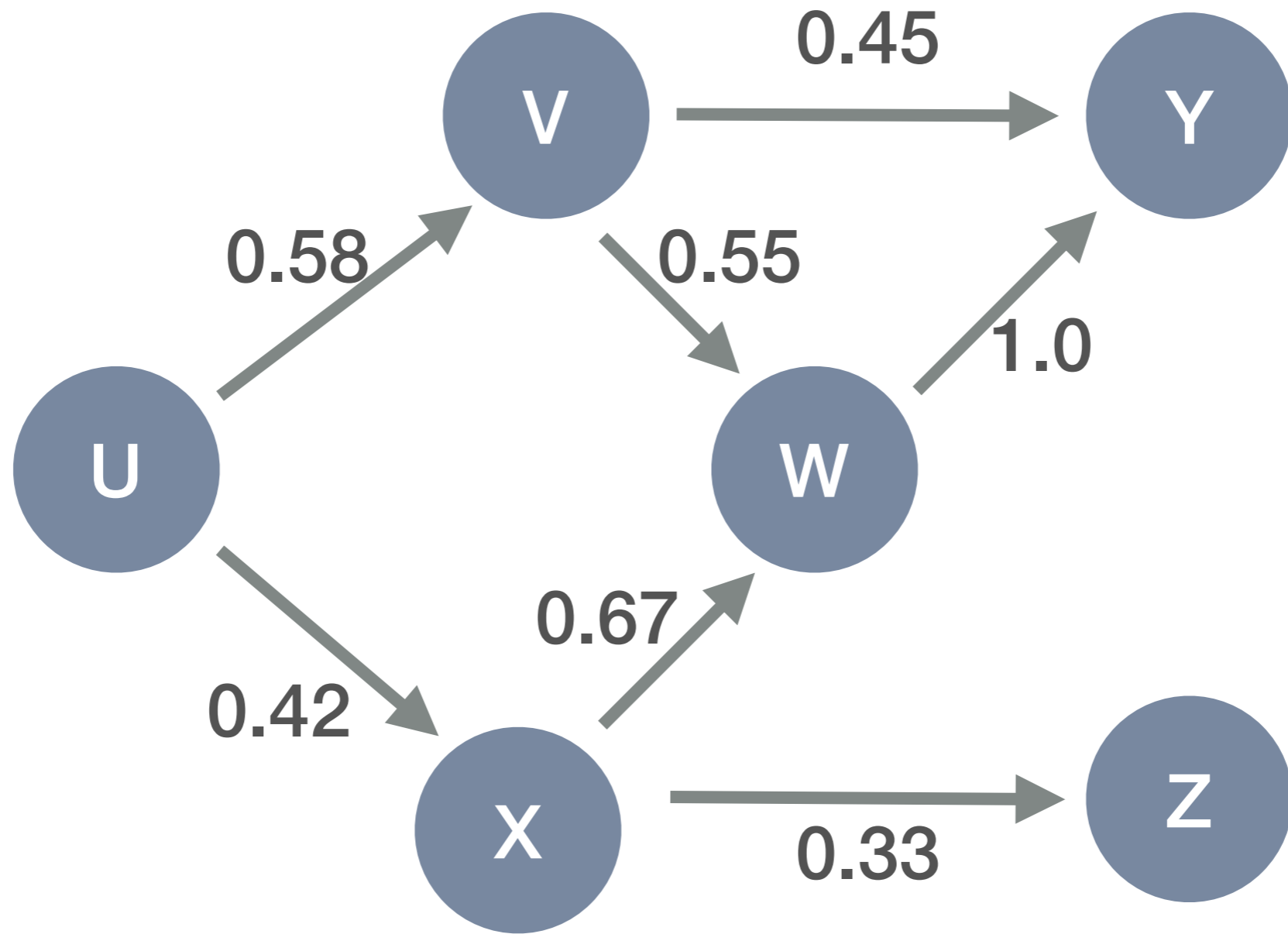




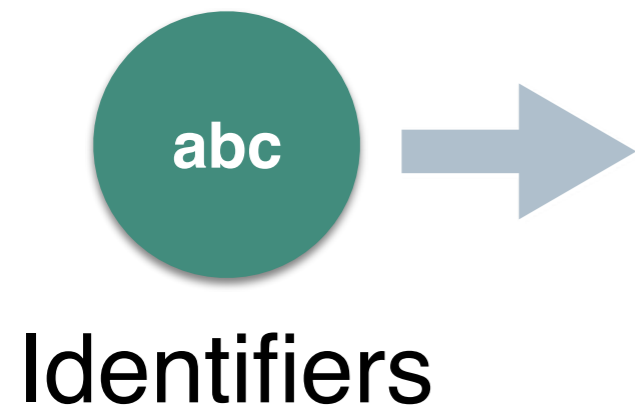




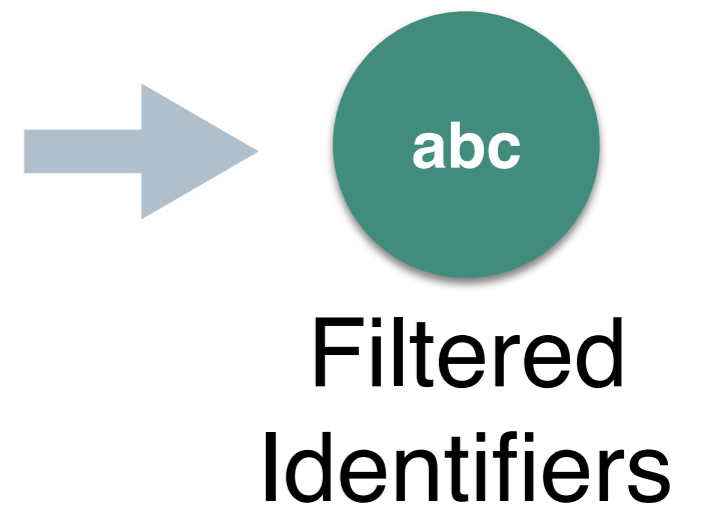




# SALLY: Filtering



- No stop-words
- More than 3 characters
- Identifiers in more than 50% of the projects
- Only tags in StackOverflow



## Tags

popular

name

new

A tag is a keyword or label that categorizes your question with other, similar questions. Using the right tags makes it easier for others to find and answer your question.

Type to find tags:

[.htaccess](#) × 44986

Directory-level configuration file used by Apache web servers.

29 asked today, 193 this week

[qt](#) × 44759

a cross-platform application development framework widely used for the development of application software that

45 asked today, 227 this week

[sql-server-2008](#) × 43956

for questions specific to the 2008 version of Microsoft's SQL Server.

18 asked today, 92 this week

[scala](#) × 43363

a general purpose programming language principally targeting the Java Virtual Machine. Designed to express common

47 asked today, 278 this week

[wcf](#) × 43207

a part of the .NET Framework that provides a unified programming model for rapidly building service-oriented applications.

14 asked today, 100 this week

[sqlite](#) × 43165

a software library that implements a self-contained, serverless, zero-configuration, transactional SQL database engine.

37 asked today, 227 this week

[function](#) × 42558

A function (also called a procedure, method, subroutine, or routine) is a portion of code intended to carry out a single,

35 asked today, 222 this week

[file](#) × 41601

A block of arbitrary information, or resource for storing information, accessible by the string-based name or path. Files are

53 asked today, 229 this week

[uitableview](#) × 40958

a class used for displaying and editing lists of information on iOS. A table view displays items in a single column.

30 asked today, 206 this week

[shell](#) × 40292

The term 'shell' refers to a general class of text-based command interpreters most often associated with the Unix & Linux

52 asked today, 240 this week

[codeigniter](#) × 40027

an open-source PHP web development framework created by EllisLab Inc and it has been adopted by British Columbia

30 asked today, 210 this week

[python-2.7](#) × 39629

the last major version in the 2.x series. This release contains many of the features that were first released in Python 3.1. Use

61 asked today, 360 this week

[validation](#) × 39374

the process of ensuring that a program operates on clean, correct and useful data.

21 asked today, 135 this week

[cordova](#) × 39029

an open-source, cross-device mobile development platform that allows developers to create mobile applications

62 asked today, 285 this week

[api](#) × 39008

An application programming interface (API) is a layer that allows software components to communicate with each other.

53 asked today, 303 this week

[actionscript-3](#) × 38710

the open source object oriented programming (OOP) language of the Adobe Flash and AIR Platforms. AS3 is

8 asked today, 38 this week

# SALLY: Generating Tags

## Filtered Identifiers

access  
accuraci  
acronym  
web  
weight  
white  
whitespac  
width  
wifi



Identifier	TF-IDF
access	0.47
accuraci	0.65
acronym	1.44
web	2.88
weight	2.57
white	0.18
whitespac	0.91
width	1.11
wifi	0.56

# SALLY: Generating Tags

## Extracted Identifiers

~~access~~  
~~acnignaci~~  
acronym  
~~wiebh~~  
whitespac  
~~acntraci~~  
whitespac  
~~acntraci~~  
whitespac  
~~acntraci~~  
whitespac  
~~acntraci~~  
white

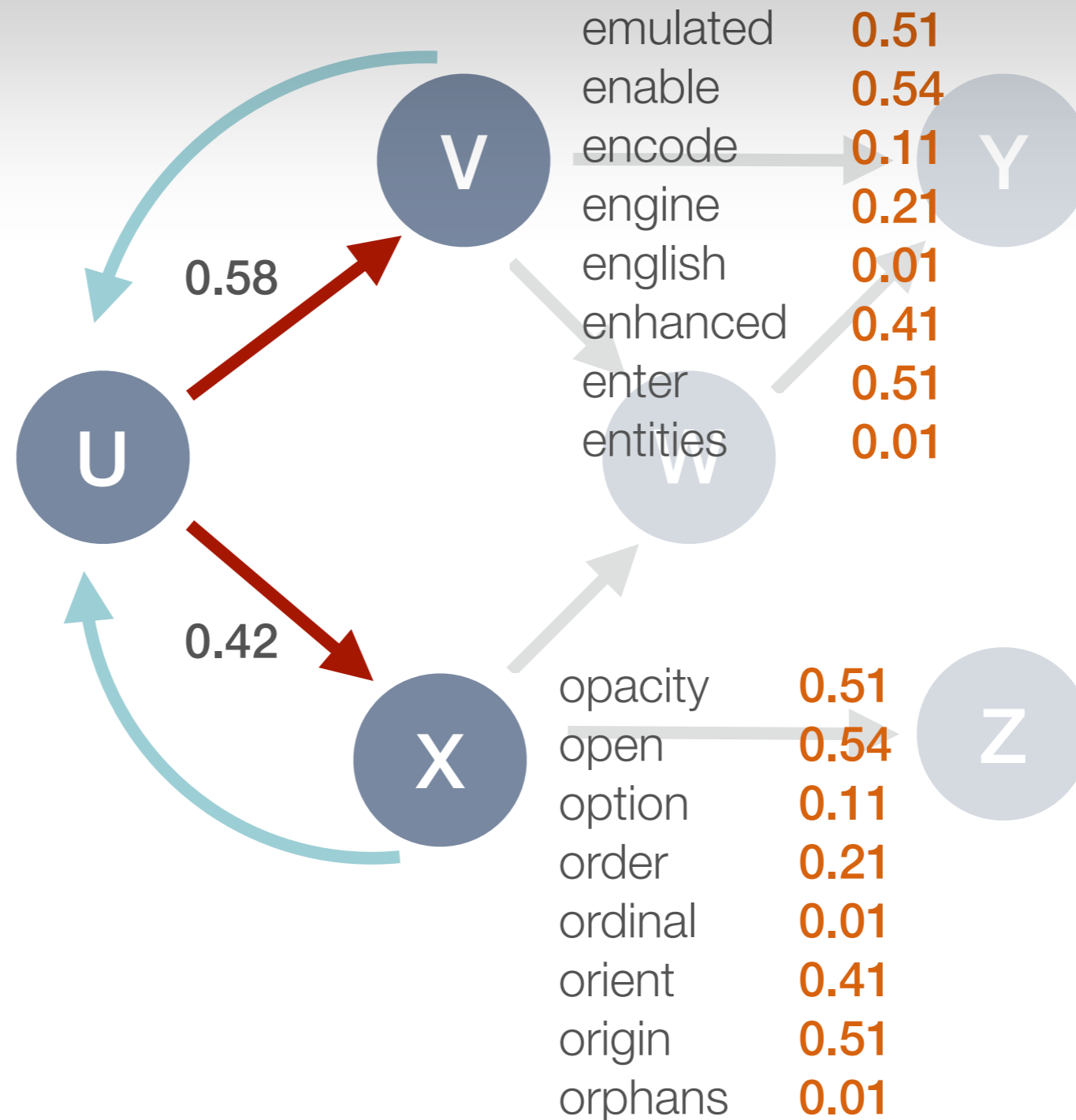


Identifier	TF-IDF	Relevanc
<del>access</del>	0.88	<del>0.07</del>
<del>acnignaci</del>	0.53	<del>0.06</del>
acronym	1.44	0.13
<del>wiebh</del>	2.88	<del>0.20</del>
<del>whitespac</del>	0.97	<del>0.08</del>
<del>acntraci</del>	0.68	<del>0.06</del>
<del>whitespac</del>	0.96	<del>0.08</del>
<del>acntraci</del>	0.47	<del>0.04</del>
<del>whitie</del>	0.58	<del>0.05</del>

# SALLY: Generating Tags

## Secondary tags

open  
enable  
emulated  
opacity  
origin  
enhanced  
...





# SALLY: Resolving Definitions

 **Primary tags**

 **Secondary tags**



WIKIPEDIA

Wiktionary

 WhatIs.com

 **stackoverflow**

[com.google.code.findbugs.findbugs-3.0.0](https://com.google.code.findbugs.findbugs-3.0.0)

**bug** 24% analysis 15% cloud 14% plugin 9% dataflow 9% annotation 8% database

5% field 5% edge 5% frame 4%

**func** 20% stylesheet 12% templates 10% xalan 10% xpath 9% apache 8% iterate

8% exslt 8% namespace 8% expr 7%



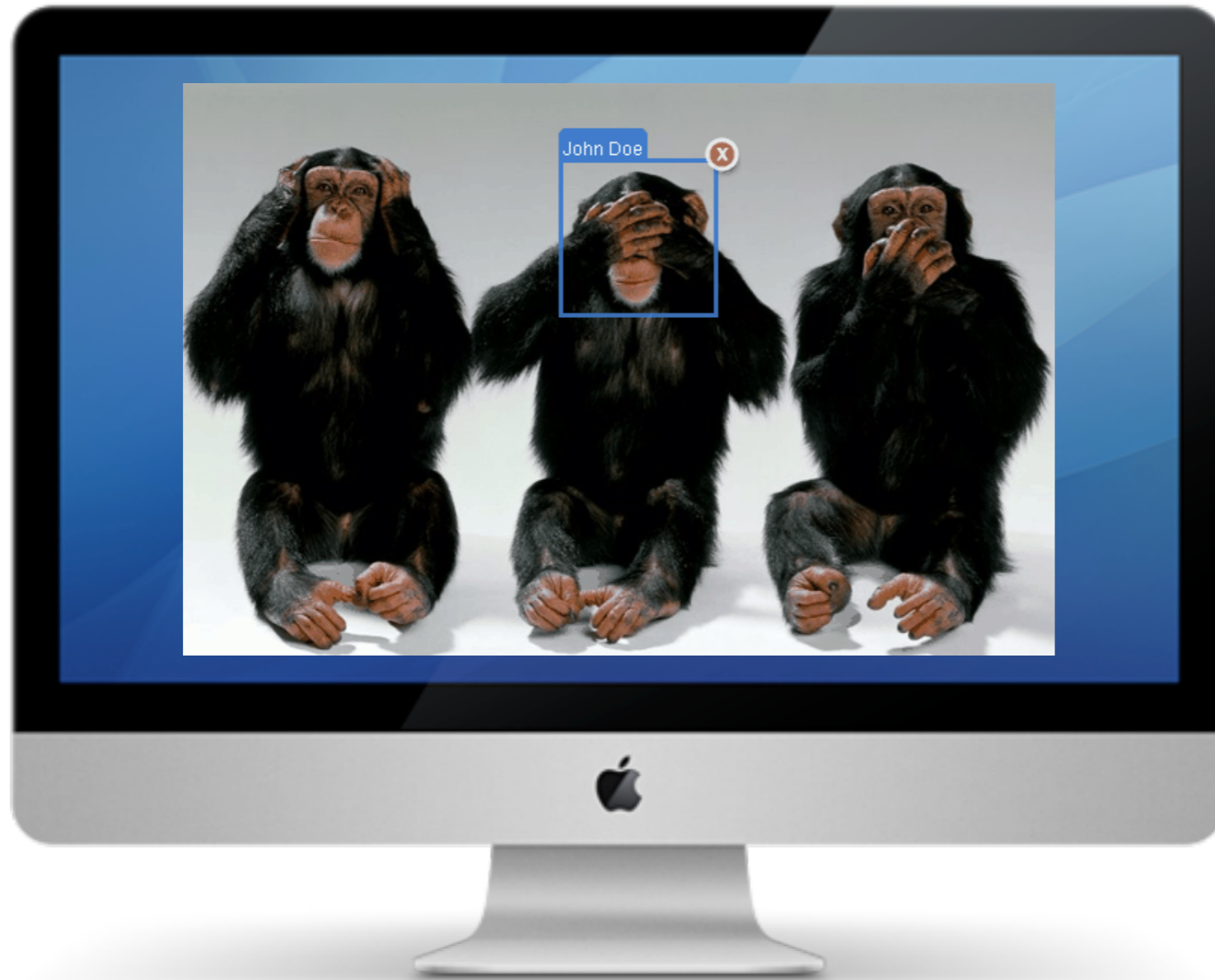
**dataflow**

[StackOverflow](#) [Wikipedia](#) [Techtarget](#)

Dataflow programming is a programming paradigm in which computations are modeled through directed graphs: nodes are instructions and data flows through the connections between them.

*Definition from StackOverflow.*

Close



# EMPIRICAL STUDY

# Experiment 1: Online tools

68

net.sourceforge  
projects

99

Dependencies



**sourceforge**

**MVNREPOSITORY**

**SALLY**

➔ **Expressiveness  
and availability**

# Experiment 2: Sally vs MUDABlue

## MUDABlue: An Automatic Categorization System for Open Source Repositories

Shinji Kawaguchi<sup>†</sup>, Pankaj K. Garg<sup>††</sup>, Makoto Matsushita<sup>†</sup> and Katsuro Inoue<sup>†</sup>

<sup>†</sup>Graduate School of Information Science and Technology, Osaka University

1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

{s-kawagt, matusita, inoue}@ist.osaka-u.ac.jp

<sup>††</sup>Zee Source

1684 Nightingale Avenue, Suite 201

Sunnyvale, California, 94807, USA

garg@zeesource.net

### Abstract

*Open Source communities typically use a software repository to archive various software projects with their source code, mailing list discussions, documentation, bug reports, and so forth. For example, SourceForge currently hosts over seventy thousand Open Source software systems. Because of the size of the rich information content, such repositories offer numerous opportunities for sharing information among projects. For example, one would like to know a set of projects that are related or similar to each other, so that the project groups can collaborate and share their work. With thousands of projects in typical repositories, however, manually locating related projects can be difficult. Hence, we propose MUDABlue, a tool that automatically categorizes software systems. MUDABlue has three major aspects: 1) it relies on no other information than the source code, 2) it determines category sets automatically, and 3) it*

of the corporations that are known to have deployed such archival service for their own internal corporate network.

For large software archives, categorizing their contents for browsing and searching is essential for effective utilization of the software archive. Automatic categorization would be helpful in several ways:

- Several *similar* software systems can be grouped together in a category for ease of browsing. For example, SourceForge categorizes software according to their primary function (editors, databases, etc.), and also has the notion of *software foundries* for related software.
- Developers working on a software system may be informed about related software. Finding related software systems has following advantages.
  1. Developers can learn “best practices” and programming idioms from existing software sys-

# Experiment 2: Sally vs MUDABlue



14 Developers



50 projects

- Online survey
- Sally vs MudaBlue tags
- 10 tags each approach
- **1-5 stars rating:  
expressiveness and  
completeness**

# Experiment 2: Sally vs MUDABlue

## **EXPRESIVENESS**

How good are the tags at describing the application domain / purpose of the software project?

## **COMPLETENESS**

Is the application domain / purpose fully described by the presented tags?

# Experiment 2: Sally vs MUDABlue

## Question 1

The following is the description for library `jtransforms-2.4.0.jar`: "JTransforms is the first, open source, multithreaded FFT library written in pure Java. Currently, four types of transforms are available: Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), Discrete Sine Transform (DST) and Discrete Hartley Transform (DHT). The code is derived from General Purpose FFT Package written by Takuya Ooura and from Java FFTPack written by Baoshe Zhang."

The official website for `jtransforms-2.4.0` is <https://sites.google.com/site/piotrwendykier/software/jtransforms>.

Please rate using a scale from 0 to 5 the expressiveness and completeness of the following sets of words. Afterwards, please give the rationale behind your decision.

SET1: `fft dct dht dst benchmark accuracy real complex thread forward`

SET2: `run push loader log column external debug thread normalization level`

Please rate the first set of words in the following aspects:\*

`fft dct dht dst benchmark accuracy real complex thread forward`

	1	2	3	4	5
Expressiveness for application domain	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Expressiveness for library purpose	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Completeness for application domain	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Completeness for library purpose	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please rate the second set of words in the following aspects:\*

`run push loader log column external debug thread normalization level`

	1	2	3	4	5
Expressiveness for application domain	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Expressiveness for library purpose	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Completeness for application domain	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Completeness for library purpose	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



**PRELIMINARY  
RESULTS**



# Example: JTransforms

*JTransforms is the first, open source, **multithreaded FFT** library written in pure Java. Currently, four types of **transforms** are available: **Discrete Fourier Transform (DFT)**, **Discrete Cosine Transform (DCT)**, **Discrete Sine Transform (DST)** and **Discrete Hartley Transform (DHT)**.*

**source**forge

MVNREPOSITORY

**S**  **LLY**

MATHEMATICS

NO TAG

FFT, DCT, DHT, DST, BENCHMARK

# Example: Bouncy castle (bcprov-jdk)

*The Bouncy Castle Java API for handling the **OpenPGP** protocol. This jar contains the **OpenPGP** API for JDK 1.5 to JDK 1.8. The APIs can be used in conjunction with a **JCE/JCA** provider such as the one provided with the Bouncy Castle **Cryptography** APIs.*

**source**forge

NOT AVAILABLE

MVNREPOSITORY

ENCRYPTION LIBRARIES

**S**  **LLY**

JCE, CERTIFICATE, CIPHER, REVOCATION, RSA

# Experiment 1: Online tools



SOURCEFORGE

**31** of transitive dependencies were **not available**  
in SourceForge

**38** projects with **overly broad tags**  
in SourceForge had specific ones in Sally

# Experiment 1: Online tools



**16%** of projects had at least one **matching tag** in both tools

**40%** of projects were tagged **only by Sally**

# Experiment 1: Online tools

## Uncategorized projects by each approach

<b>SourceForge</b>	<b>MVN Repo</b>		<b>Sally</b>	
<b>Categories</b>	<b>Categories</b>	<b>Tags</b>	<b>Primary</b>	<b>Secondary</b>
<b>23</b>	<b>93</b>	<b>67</b>	<b>2</b>	<b>71</b>
<b>21.69%</b>	<b>56.36%</b>	<b>40.6%</b>	<b>1.2%</b>	<b>42.51%</b>

# Experiment 2: Sally vs MUDABlue

*HtmlUnit is a "GUI-Less browser for Java programs". It **models HTML documents** and provides an API that allows you to invoke pages, fill out forms, click links, etc... just like you do in your "normal" browser."*

Sally	
html	web
svg	jsx
dom	dom
border	functional
script	func

MUDABlue	
entries	external
debug	component
escape	loader
control	level
extensions	log

# Experiment 2: Sally vs MUDABlue

## Expressiveness

SALLY

MUDABlue

5 ★

16.8%

1.1%

4 ★

32.1%

11.4%

3 ★

27.5%

16.1%

2 ★

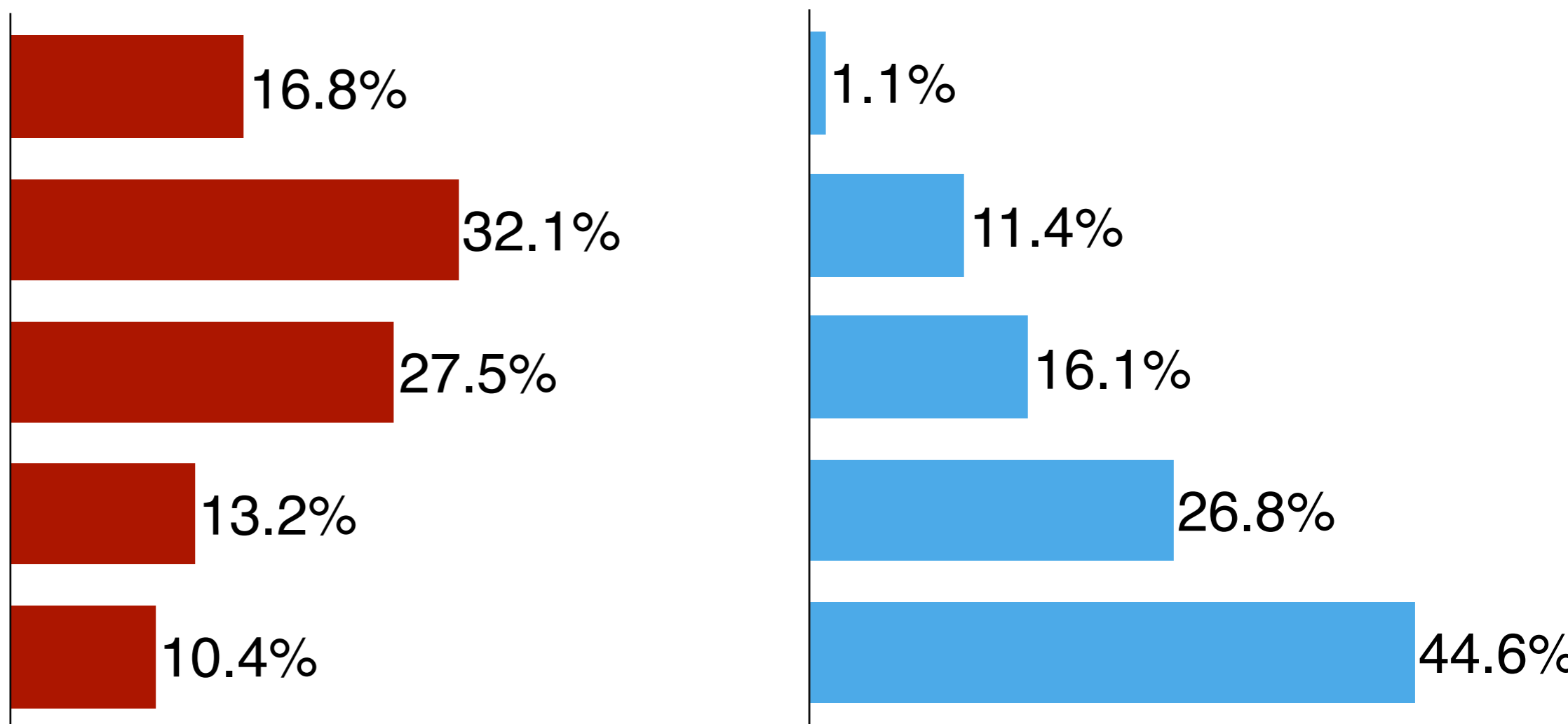
13.2%

26.8%

1 ★

10.4%

44.6%



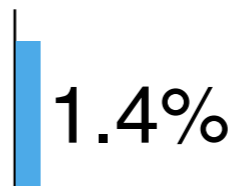
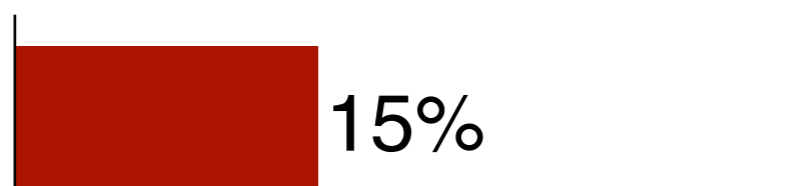
# Experiment 2: Sally vs MUDABlue

## Completeness

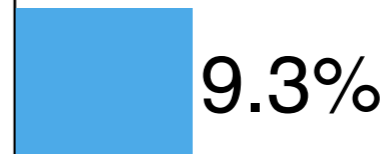
SALLY

MUDABlue

5 ★



4 ★



3 ★



2 ★



1 ★







Home

Projects

Categories

## Welcome to Sally!

**Sally** is an automatic categorization engine for Maven libraries. It is the result of the work made in the MSc thesis *Automatic Multi-label Categorization of Java Applications Using Dependency Graphs* by Santiago V. Baldrich from Universidad Nacional de Colombia in Bogotá, Colombia.

In the projects tab you can find a list of Maven projects, you can use the filters on the top of the page to narrow down the list to show the projects with a particular *groupId* or *artifactId*. If you click on any item from the list, you'll be presented with the categories **Sally** obtained for that particular project. If you want to find the definition of any category just click on it and you'll be presented with its definition taken from well-known sites if available.

The categories are shown in two colors. The **blue** ones correspond to the primary categories of the project, i.e. the categories obtained from analyzing it by itself and not taking into account its dependencies. On the other hand, **red** categories arise from the analysis of the dependencies of the project. Therefore, projects with no dependencies would not display any red categories.

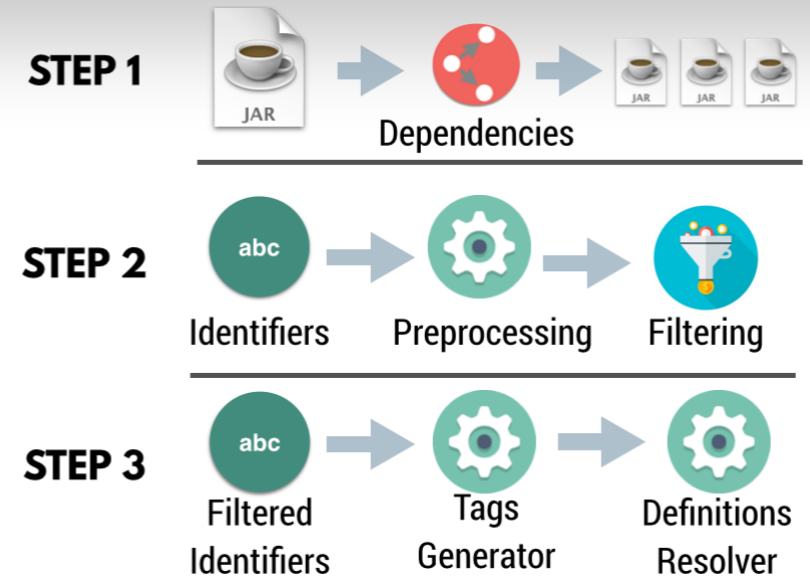
<http://sally.meteor.com>

<http://www.cs.wm.edu/semeru/sally/>

## Current Issues

1. Source code is not always available
2. Predefined categories may not be sufficient
3. Supervised categorization is not always possible

## SALLY



## Experiment 1: Online tools

SourceForge	MVN Repo		Sally	
Categories	Categories	Tags	Primary	Secondary
23	93	67	2	71
21.69%	56.36%	40.6%	1.2%	42.51%

Uncategorized projects by approach

## Experiment 2: Sally vs MUDABlue

*HtmlUnit is a "GUI-Less browser for Java programs". It models HTML documents and provides an API that allows you to invoke pages, fill out forms, click links, etc... just like you do in your "normal" browser."*

Sally		MUDABlue	
html	web	entries	external
svg	jsx	debug	component
dom	dom	escape	loader
border	functional	control	level
script	func	extensions	log