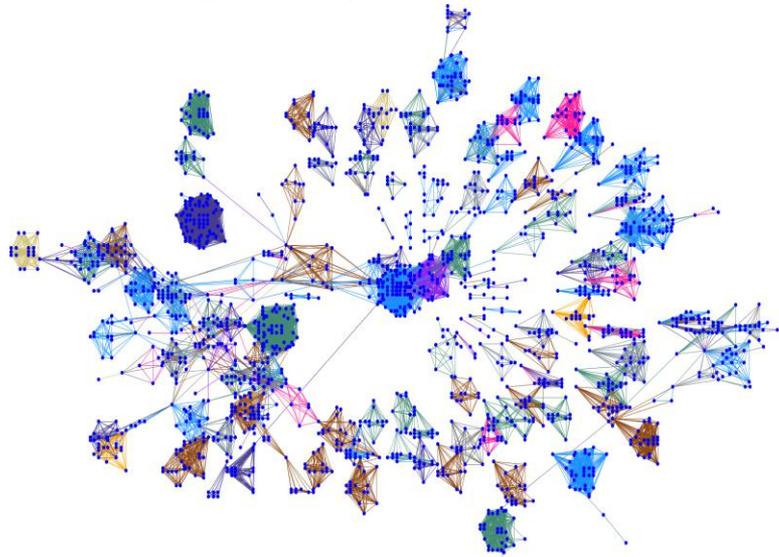




galaxyadvisors



Getting Started With Condor

Contents

- **Getting Started**
- Collecting Web Content
- OneDegreeCollector
- Building your own Startlists
- Collecting your E-Mail
- Collecting Facebook Data
- Collecting Wikipedia Data
- Collecting CoolPeople
- Coolhunting Blueprint

Data
Visualization
&
Analysis

Web Portal

Web Analysis &
Visualization
People & Concept Network
Visualization
Trend Curves
Information Spheres

Condor

In-depth Analysis
Dynamic Visualization
People & Concept Network
Analysis

3rd Party

R statistics
SPSS, Matlab
Excel
Ucinet, Pajek SNA
Gephi graph viz

Data
Processing
&
Filtering

CondorCore

Prediction

Regression
Neural Network
Genetic Algorithm
Kalman Filter
SVM

Sentiment

Positivity
Negativity
Dynamic BoW

SNA

Betweenness
Degree
Contribution
Index
Graph Layout

IR

Lucene
TF-IDF
BNC
Proper Noun
Information
Templates

Reputation

Decay
Temporalize
DoS
MVP

Creativity

COIN detection
Cliques
Core/Periphery
ART

Export

Excel
MySQL
PDF
Txt

Data
Collection

Wikipedia

Knowledge

Google/Yahoo/Bing

Crowds

Twitter

Crowds

News/Scholar/Medline

Experts

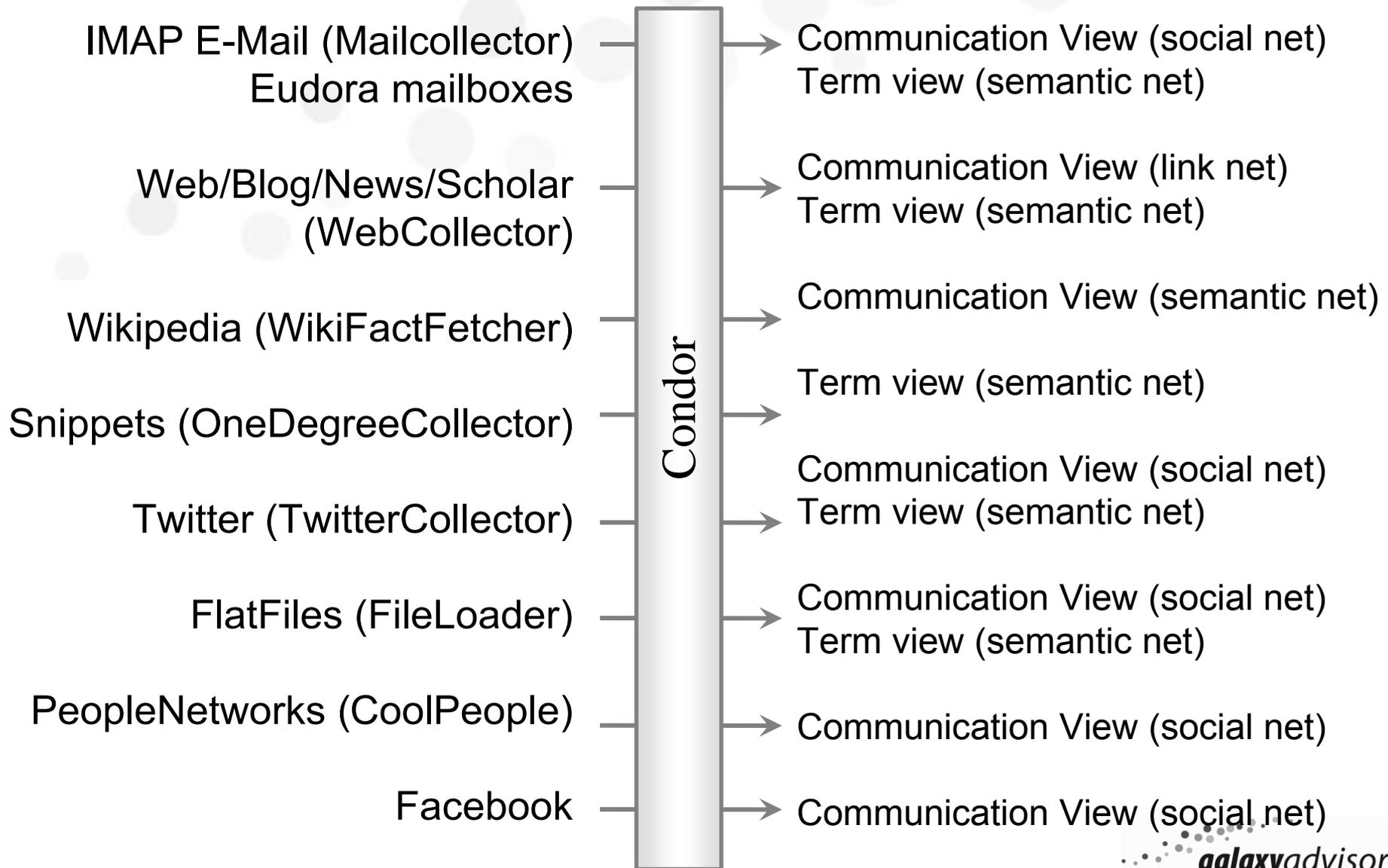
Forums/Blogs

Swarm

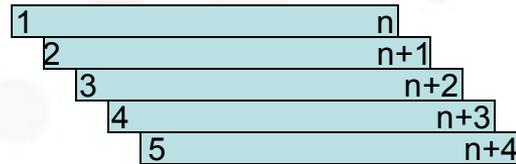
MailCollector

Swarm

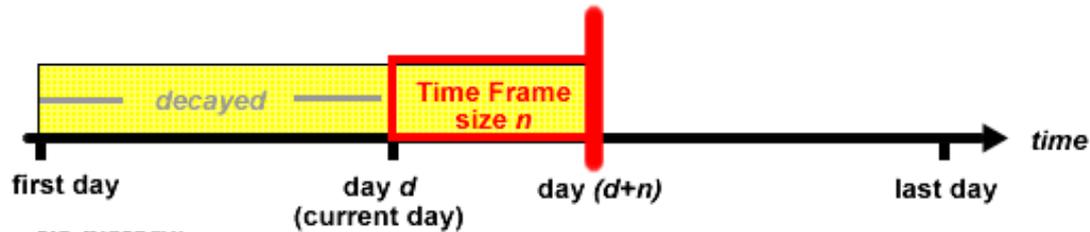
Getting Data into Condor



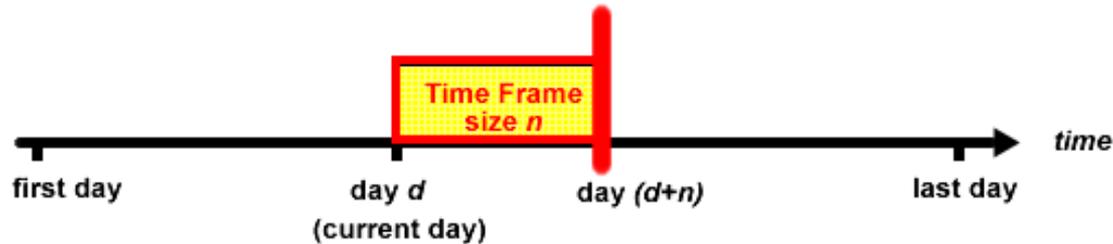
Temporal Visualization by a Sliding Time Frame



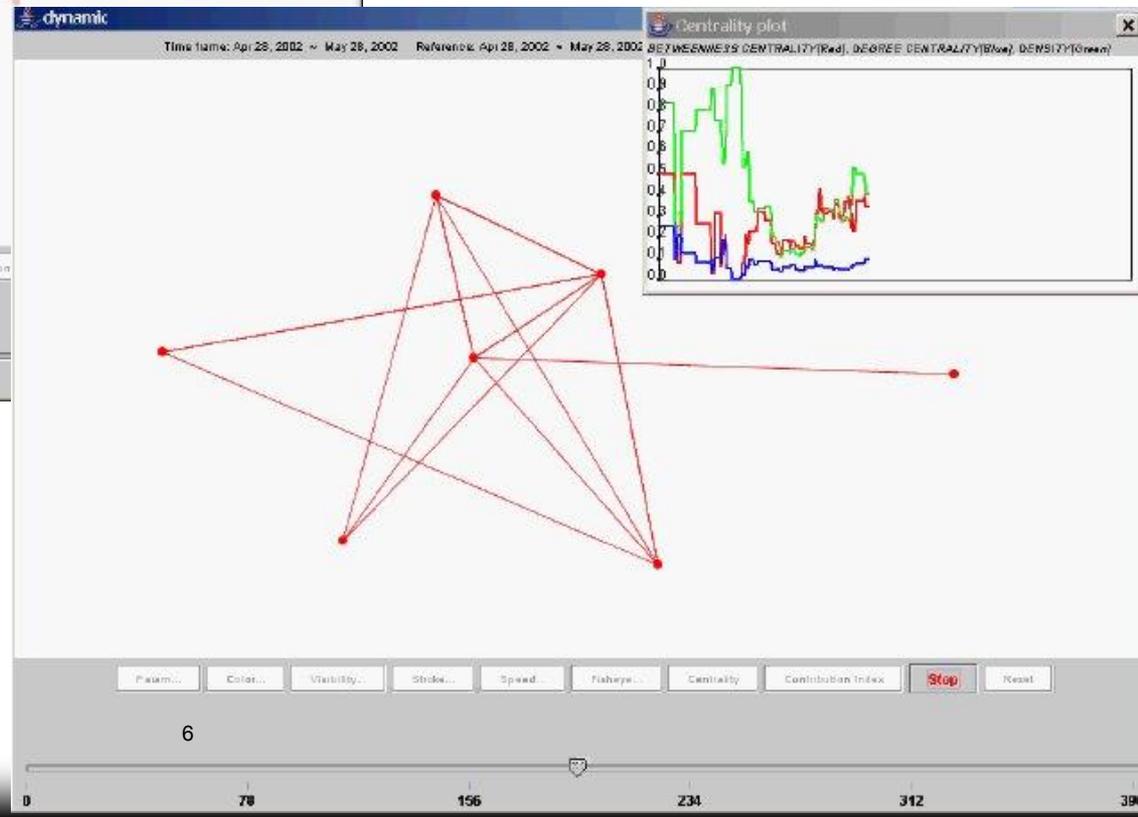
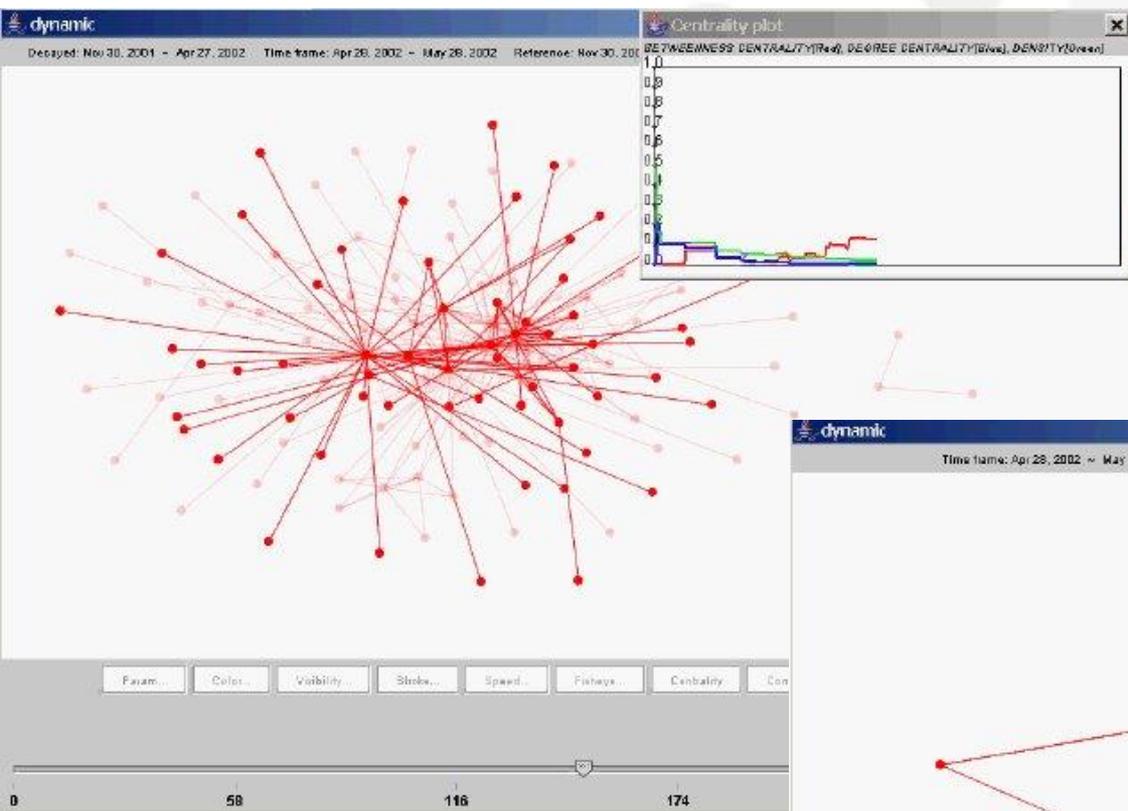
With history:



no history:



With and without history



Preparation

- Install MySQL
- Install Java (only Windows)
- Install Java 3D (only Windows)
- Start Java (if it does not run yet)

Contents

- Getting Started
- **Collecting Web Content**
- OneDegreeCollector
- Building your own Startlists
- Collecting your E-Mail
- Collecting Facebook Data
- Collecting Wikipedia Data
- Collecting CoolPeople
- Coolhunting Blueprint

Collect Web Content

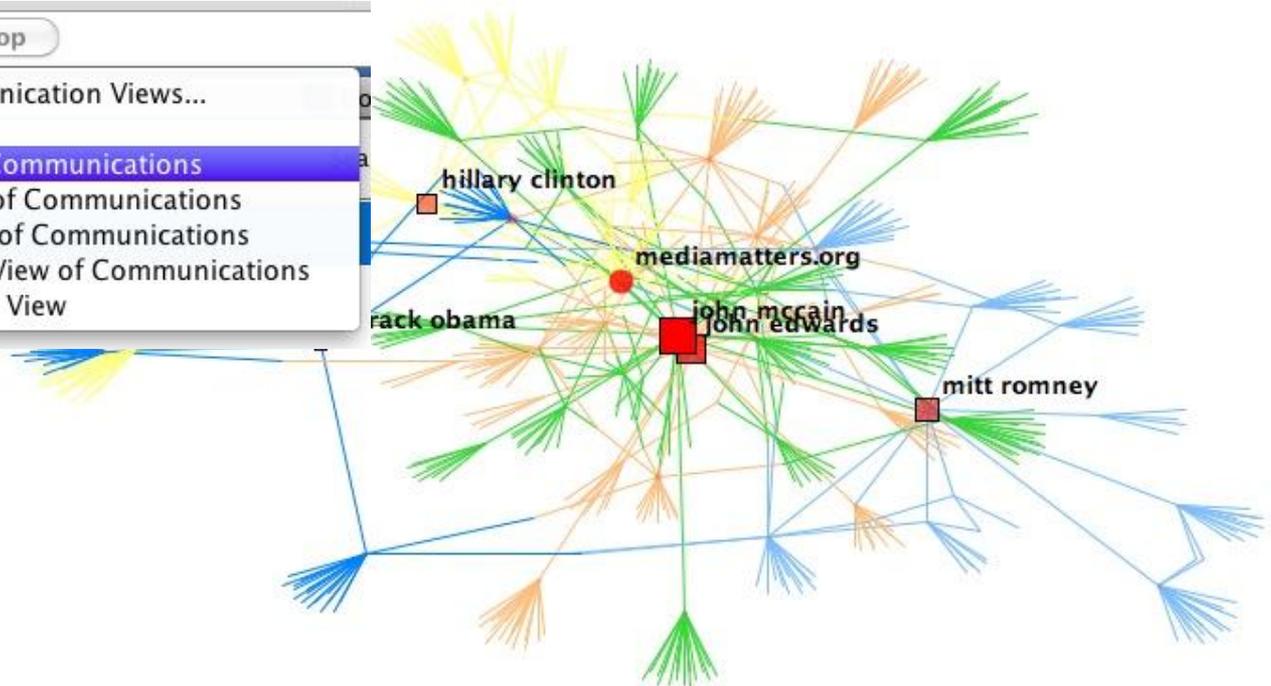
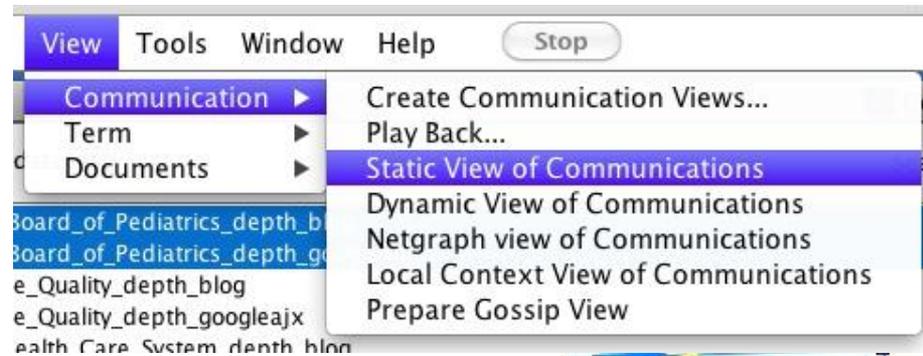
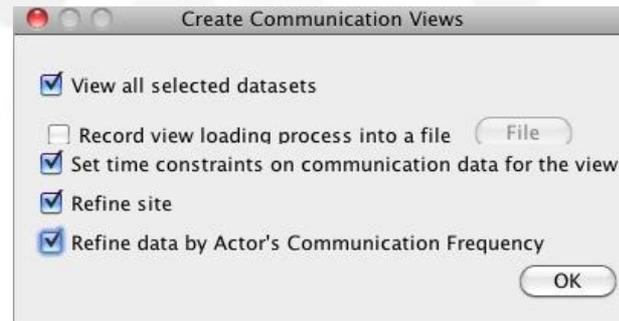
The screenshot displays the 'Web Collector' application interface. At the top, a menu bar includes 'File', 'Edit', 'View', 'Tools', 'Window', and 'Help'. The 'Tools' menu is open, listing options such as 'Web Collector...', 'One Degree Collector...', 'Mail Collector...', 'WikiFactFetcher...', 'Twitter Collector...', 'FaceBookFriends...', 'File Loader...', and 'Cool People...'. Below the menu, the 'Available datasets' section is empty.

The main interface is divided into several sections:

- Database Help:** Includes an 'API' section with radio buttons for search engines: 'use Google Search (G)', 'use Google Blog Search', 'use MSN Live WebSearch', 'use Google News Search', and 'use Google Scholar Search'.
- Query:** A 'Text Query' section with instructions: '- for exact phrase search, enclose queries with "..." (e.g: "coca cola", "virgin cola")' and '- separate multiple queries with commas (e.g.: coca cola, virgin cola)'. Below it is a 'URL Query' section with instructions: '- enter the websites you want to search (e.g.: http://www.cocacola.com, virgincola.com, pepsi.com/news.html)' and '- separate multiple queries with commas'.
- Optional:** A section for date restriction: 'restrict by date. Start: 04/25/2011' and 'End: 04/25/2011'.
- Options:** Includes 'Results of Top' (set to 20), 'Degree of Separation' (set to 2), a checkbox for 'Collect content (this takes more time)', and a checked checkbox for 'Preview results'.
- Database Name:** A small dialog box titled 'Create New Database' with a text input field for the database name.
- Results:** A list of search results for the query 'barack obama'. The first result is 'Preview results for query barack'. The list includes various URLs and timestamps, such as 'http://www.thecarpetbaggerreport.com/archives/11513.html' and 'http://www.thinkonthesethings.wordpress.com/2007/07/19/how-a-Democrat.Anyone-who-has-read-The-Audacity-of-Hope-knows-this-He-has-also-given'.

At the bottom left, a red message states: 'No database connection, use the menu of this form to open or create one before starting the collector.' At the bottom right, there are 'Close' and 'Start' buttons.

Communication View



Term view index

Term processing

Use **existing index** (uncheck to create a new index)

Terms Selection Strategy

Preview terms

Upload list of terms

Indexing

Include **numbers** (whether numbers should be indexed)

Use Porter **stemming** (whether words should be reduced to their stem)

Set constraints on document date

Stopwords

Stopwords are highly frequent words like "the" or "for". Stopwords are skipped during...

use **minimal** stopwords list

use **default** stopwords list

use **extended** stopwords list

use **blog** stopwords list

use **custom** stopwords list

(one word or phrase per line, no commas, lowercase words)

Term view index - 2

Term Selection

This dialog lets you select which terms will be used for the term view
Click the words or phrases from the left column to add it to the right selection

Add Words
word (*norm. frequency*) (click to add)

- v6.00.2800.1165 (56.0)
- mimeole (56.0)
- priority (56.0)
- microsoft (56.0)
- msmail (56.0)
- charset (32.0)
- iso-8859-1 (32.0)
- http (29.0)

Add 10 from the top

Add Phrases
phrase (*norm. frequency*) (click to add) -->

- microsoft before (56.0)
- microsoft very (56.0)
- msmail before (56.0)
- msmail very (56.0)
- charset very (32.0)
- charset before (32.0)
- http before (29.0)
- http very (29.0)

Add 10 from the top

Add your own term

Selected Terms
(click item to remove)

- produced (56.0)
- survey (36.0)
- thomas (31.0)
- mobile (30.0)
- deloitte (28.0)
- nico (28.0)
- ckn (28.0)
- consulting (28.0)
- sent (26.0)
- schmalberger (26.0)
- web (26.0)
- please (26.0)
- sent before (26.0)
- sent very (26.0)
- web before (26.0)
- web very (26.0)
- before mailing (26.0)
- very mailing (26.0)
- 2002 (25.0)
- thurgauerstrasse (24.)
- converted (24.)
- pgloor (24.)
- original (24.)
- zurich (24.)
- users (24.)
- boundary (24.)
- people (23.0)

sort terms by **norm. frequency**

sort terms by **name**

Factor actor's weight
in when computing communication weight

Index dates
(necessary for dynamic termview)

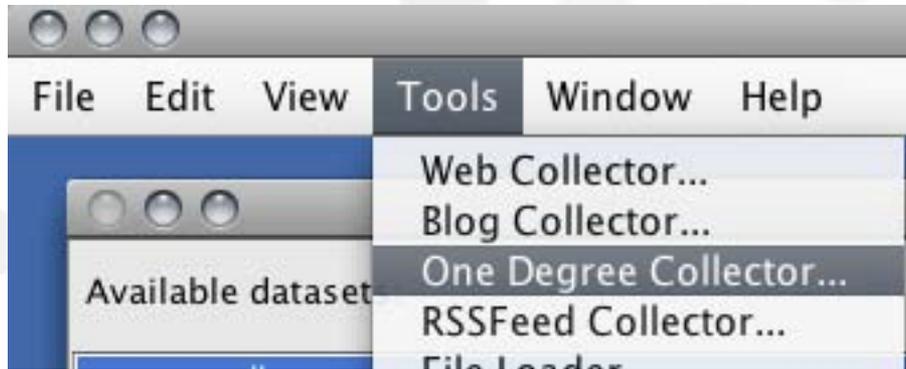
Keep top % of cooc.

All your previous terms
selections are automatically stored.

Contents

- Getting Started
- Collecting Web Content
- **OneDegreeCollector**
- Building your own Startlists
- Collecting your E-Mail
- Collecting Facebook Data
- Collecting Wikipedia Data
- Collecting CoolPeople
- Coolhunting Blueprint

One-Degree-Collector



- Complementary to the Blog Collector
- Fetches only one degree
- Retrieved websites are not aggregate

One-Degree-Collector - UI

One Degree Collector

Database Help

API

- use Google Blog Search
- use MSN Live WebSearch
- use Google News Search
- use Google Scholar Search
- use Google Search (G) (limited number of results!)
- use Yahoo Search

Query

Text Query

- for exact phrase search, enclose queries with "..." (e.g: "coca cola", "virgin cola")
- separate multiple queries with commas (e.g.: coca cola, virgin cola)

URL Query

- enter the websites you want to search (e.g.: http://www.cocacola.com, virgincola.com, pepsi.com/news.html)
- separate multiple queries with commas

Optional: restrict by date. Start: 05/21/2009 End: 05/21/2009 Number of intervals: 1

Options

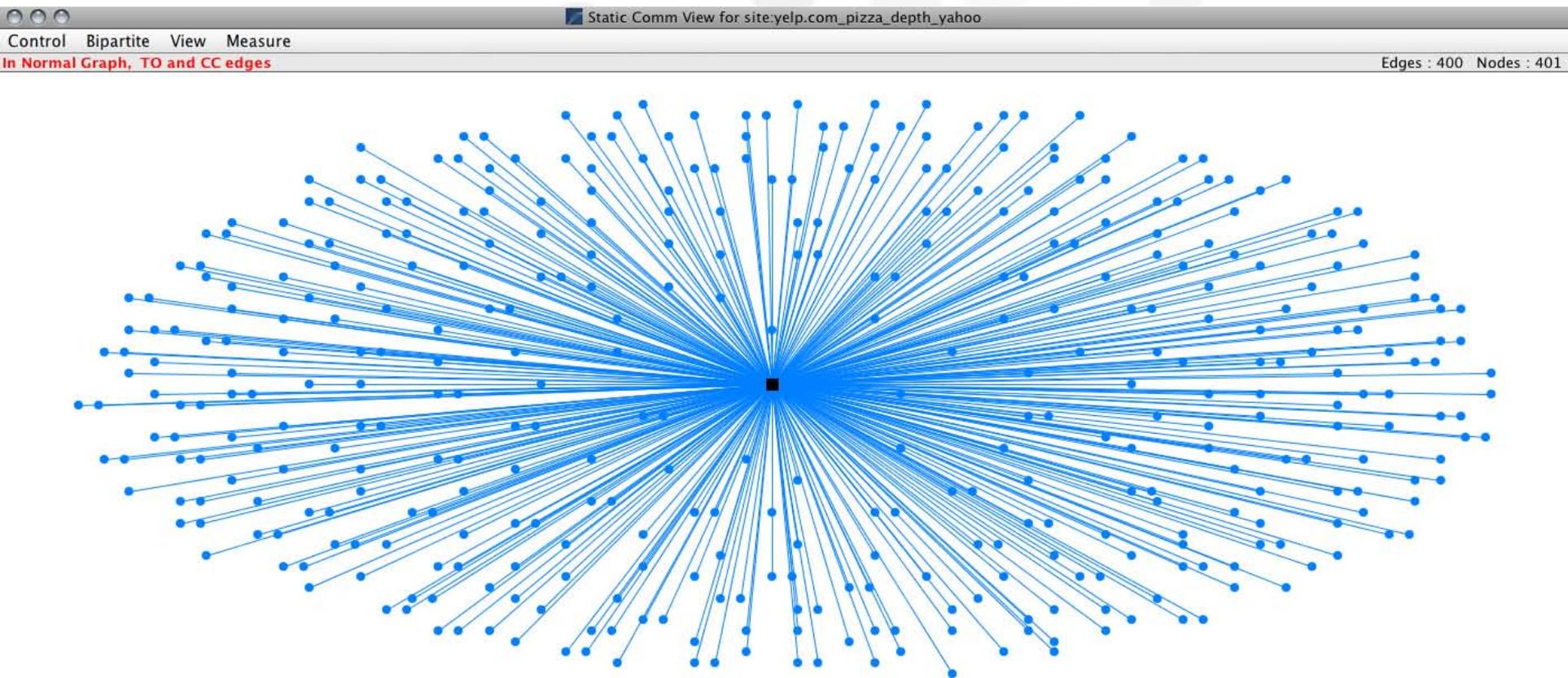
Results of Top 20 Collect content (this takes more time)

Connected to database "enron"

Close Start

- GUI resembles Blog Collector

One-Degree-Collector - result



- typical result of one-degree search

Contents

- Getting Started
- Collecting Web Content
- OneDegreeCollector
- **Building your own Startlists**
- Collecting your E-Mail
- Collecting Facebook Data
- Collecting Wikipedia Data
- Collecting CoolPeople
- Coolhunting Blueprint

Creating Term View Without OneDegreeCollector Start List: Create Stoplist First

Add Words
word (norm. frequency) (click to add)

deutschland (44.0)
deutschen (41.0)
allgemein (36.0)
unternehmen (36.0)
veröffentlicht (35.0)
soll (35.0)
euro (35.0)
geld (35.0)

Add 10 from the top

Add Phrases
phrase (norm. frequency) (click to add)

mit dem (38.0)
und der (37.0)
mit einem (34.0)
auf dem (30.0)
mit einer (29.0)
nicht mehr (29.0)
der deutschen (27.0)
sie sich (26.0)

Add 10 from the top

Add your own term

Selected Terms
(click item to remove)

januar (41.0)
abonnieren (41.0)
neuen (41.0)
ohn (41.0)
heut (40.0)
diesen (40.0)
februar (40.0)
medien (39.0)
kontakt (39.0)
august (38.0)
denn (38.0)
mich (38.0)
mal (38.0)
etwa (38.0)
juli (37.0)
diesem (37.0)
sowi (37.0)
hat (36.)
mail (36.)
die (36.)
sehr (36.0)
kategorien (36.0)
googl (36.0)
dezemb (36.0)
erst (36.0)
machen (36.0)
recht (36.0)

Word Cloud

Settings

- sort terms by norm. frequency
- sort terms by name
- Factor actor's weight in when computing communication v
- Index dates (necessary for dynamic termview)

Keep top % of cooc.

All your previous terms selections are automatically stored.

Use previous selections

Only Export selected terms

OK, collect communications ...

... then use this stop list for the term view

Indexing

- Include **numbers** (whether numbers should be indexed)
- Use Porter **stemming** (whether words should be reduced to their stem)
- Set constraints on document date

Stopwords

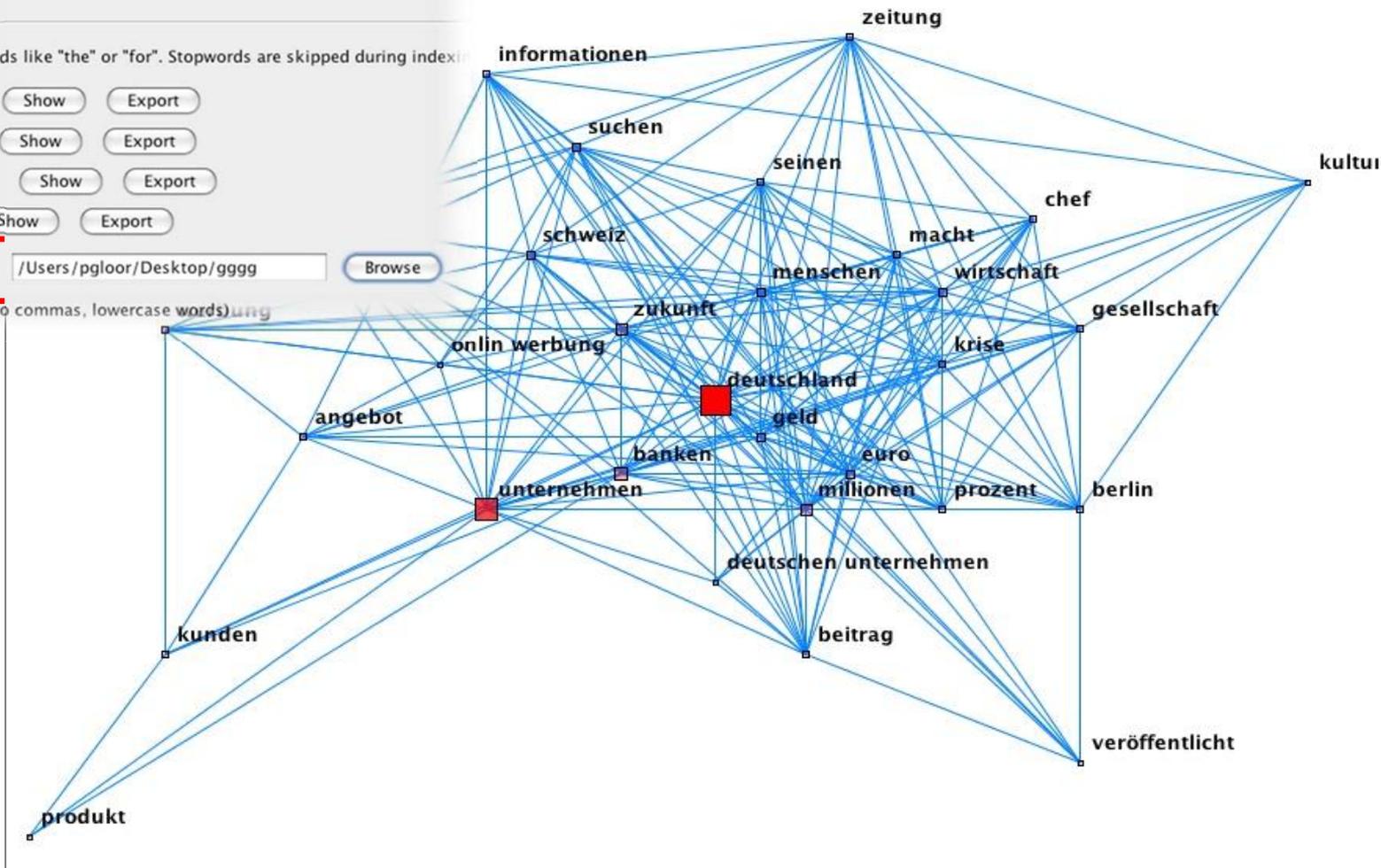
Stopwords are highly frequent words like "the" or "for". Stopwords are skipped during indexing.

- use **minimal** stopwords list
- use **default** stopwords list
- use **extended** stopwords list
- use **blog** stopwords list
- use **custom** stopwords list

(one word or phrase per line, no commas, lowercase words)

Term View for private_banking_kunde_depth_blog

Edges : 217 Nodes : 30

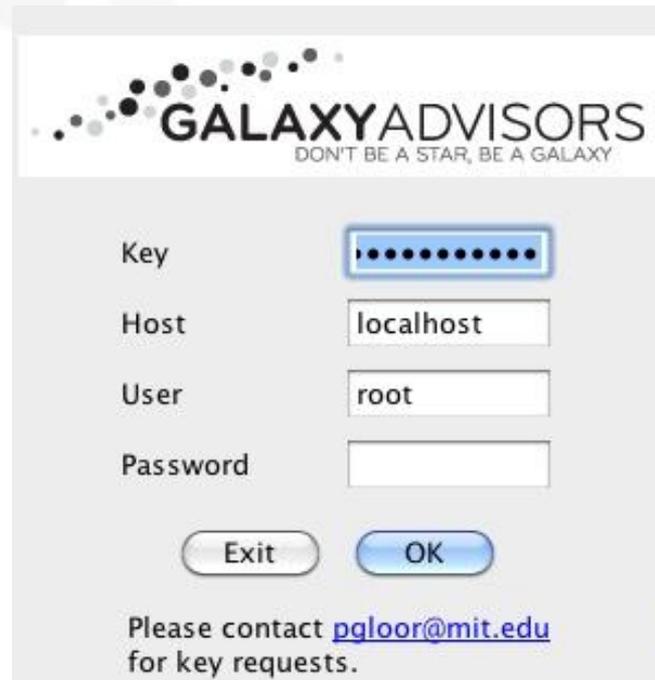


Contents

- Getting Started
- Collecting Web Content
- OneDegreeCollector
- Building your own Startlists
- **Collecting your E-Mail**
- Collecting Facebook Data
- Collecting Wikipedia Data
- Collecting CoolPeople
- Coolhunting Blueprint

Collect E-Mail

- `java -Xmx2048M -jar condor-2.1.jar`

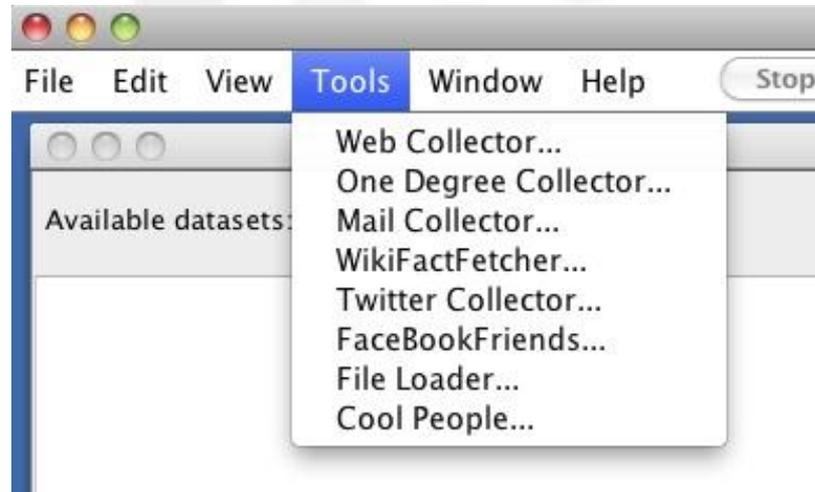


The screenshot shows a dialog box titled "GALAXYADVISORS" with the tagline "DON'T BE A STAR, BE A GALAXY". It contains four input fields: "Key" (masked with dots), "Host" (containing "localhost"), "User" (containing "root"), and "Password" (empty). There are "Exit" and "OK" buttons at the bottom. A footer note says "Please contact pgloor@mit.edu for key requests."

Condor Key

MySQL password
(default: no password)

Tools to collect data



Left side: enter here the specification of the mailbox

Right side: database related data, eg. username: root, no password

MailCollector

The screenshot shows the 'Galaxy Mail Collector' application window. It is divided into two main sections: 'Mailbox' on the left and 'Database' on the right. The 'Mailbox' section includes input fields for Username, Password, Host, Folder (set to 'ALL'), and Port (set to '993'). Below these are radio button options for Content (yes/no), SSL (no/yes), and Protocol (pop/imap). A 'Choose folder' button is located next to the Folder field. The 'Database' section includes input fields for Username, Password, Host (set to 'localhost'), Database Name (set to 'DBMail_2010_06_30212723'), and DataSet Name (set to 'myDs'). Below these are radio button options for Anonymize (yes/no), Format (Core/Condor), and Clear database (yes/no). At the bottom of the window are 'Close' and 'Start' buttons.

For username, host, port, and ssl check with your email provider (for gmail, see next slide)

Content: yes will download the whole emails, w/o content only the sender, recipients and the subject line are downloaded

Here you can choose specific folders to download

Anonymize will replace email addresses with random identifiers

Delete the present data in the database?

Settings for gmail

Galaxy Mail Collector

Galaxy Mail Collector

Mailbox

Username: yourname@gmail.com

Password: Your gmail password

Host: imap.gmail.com

Folder: ALL

Port: 993

Content: yes no

SSL: no yes

Protocol: pop imap

Database

Username:

Password:

Host: localhost

Database Name: DBMail_2010_06_30212723

DataSet Name: myDs

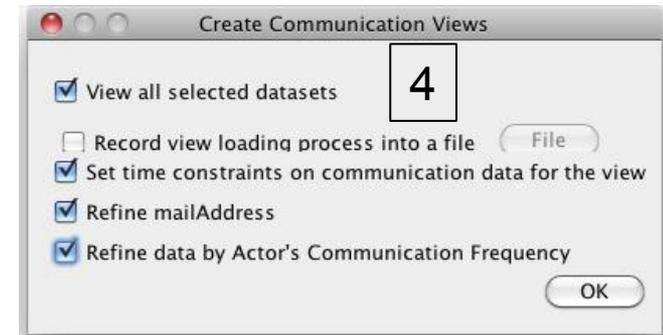
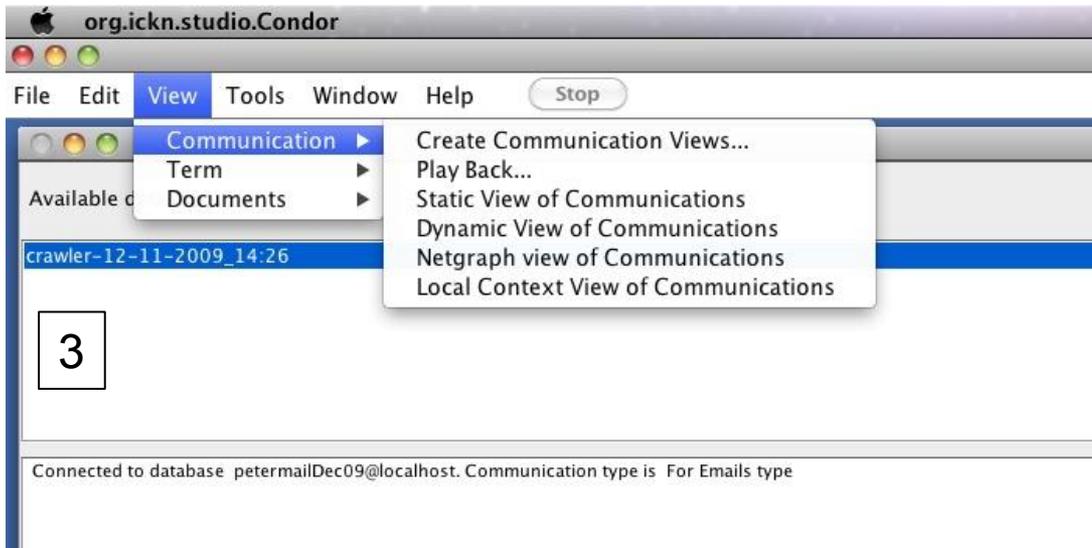
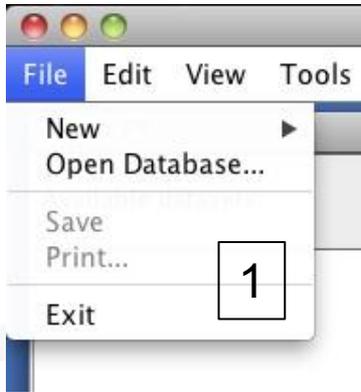
Anonymize: yes no

Format: Core Condor

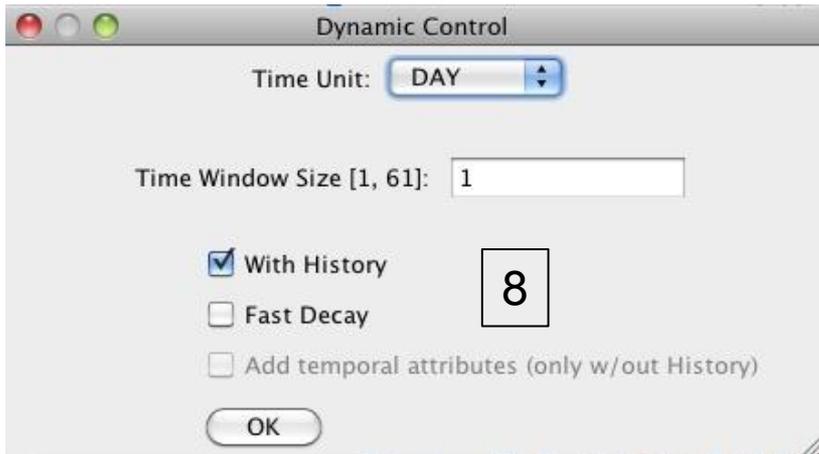
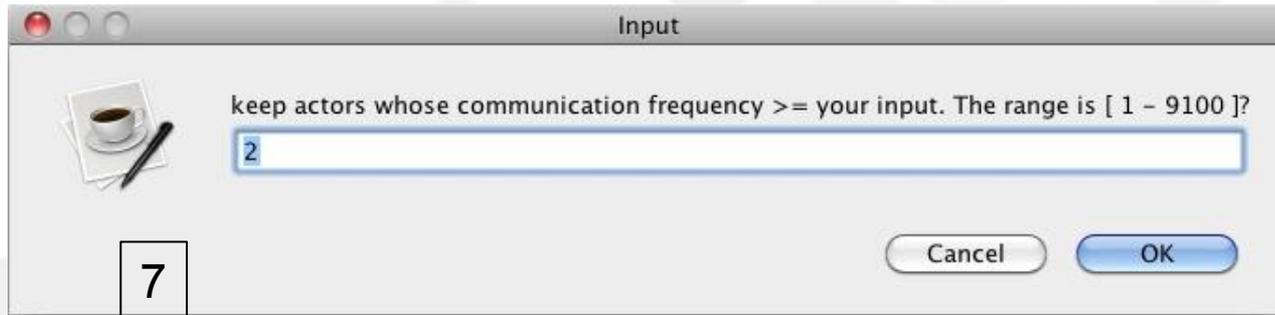
Clear database: yes no

Don't forget the access information for your mysql database on the right, then press start. It might take a while (esp. with huge mailboxes) before you see a progress bar.

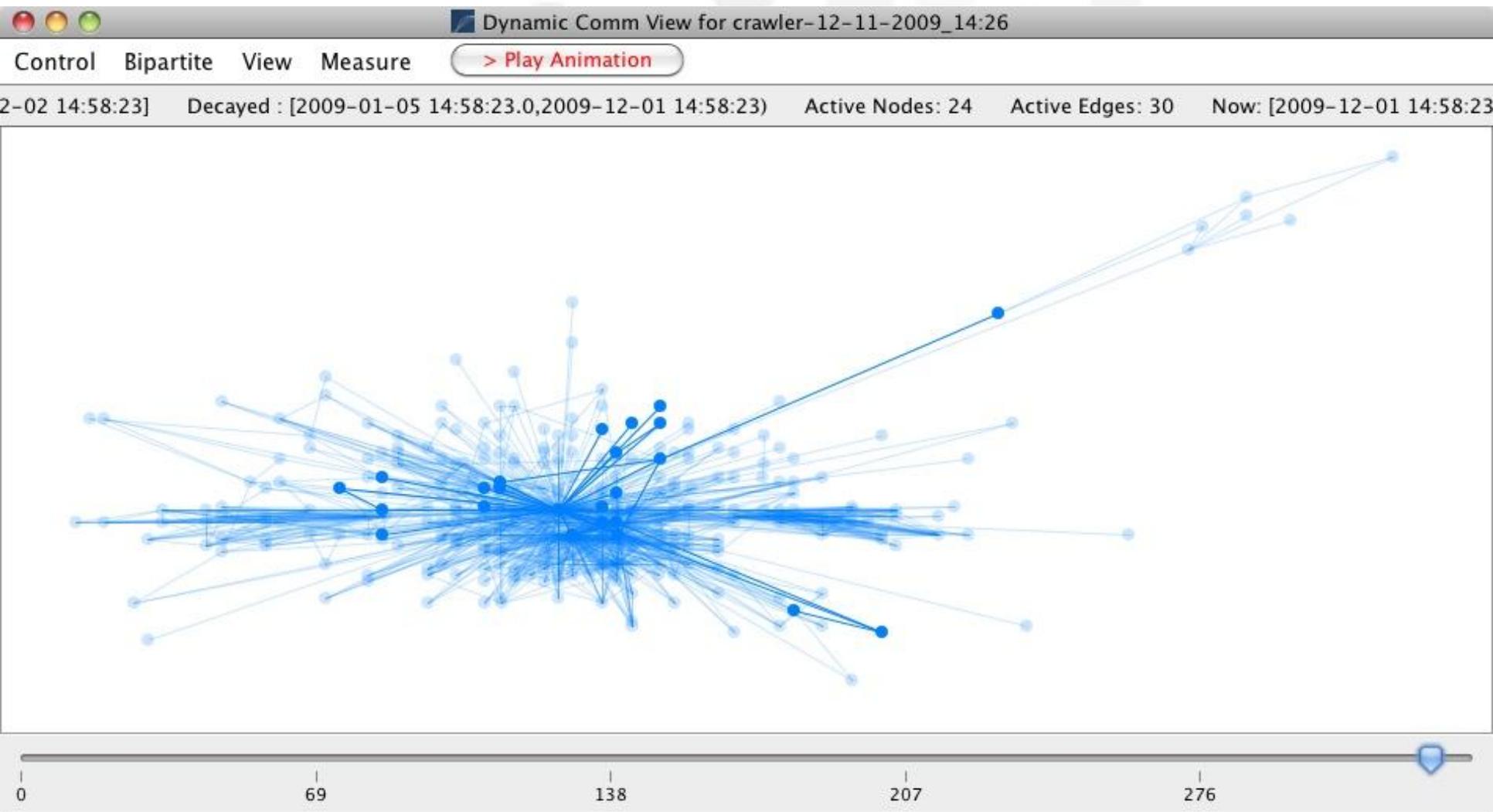
Visualize Mail-Data



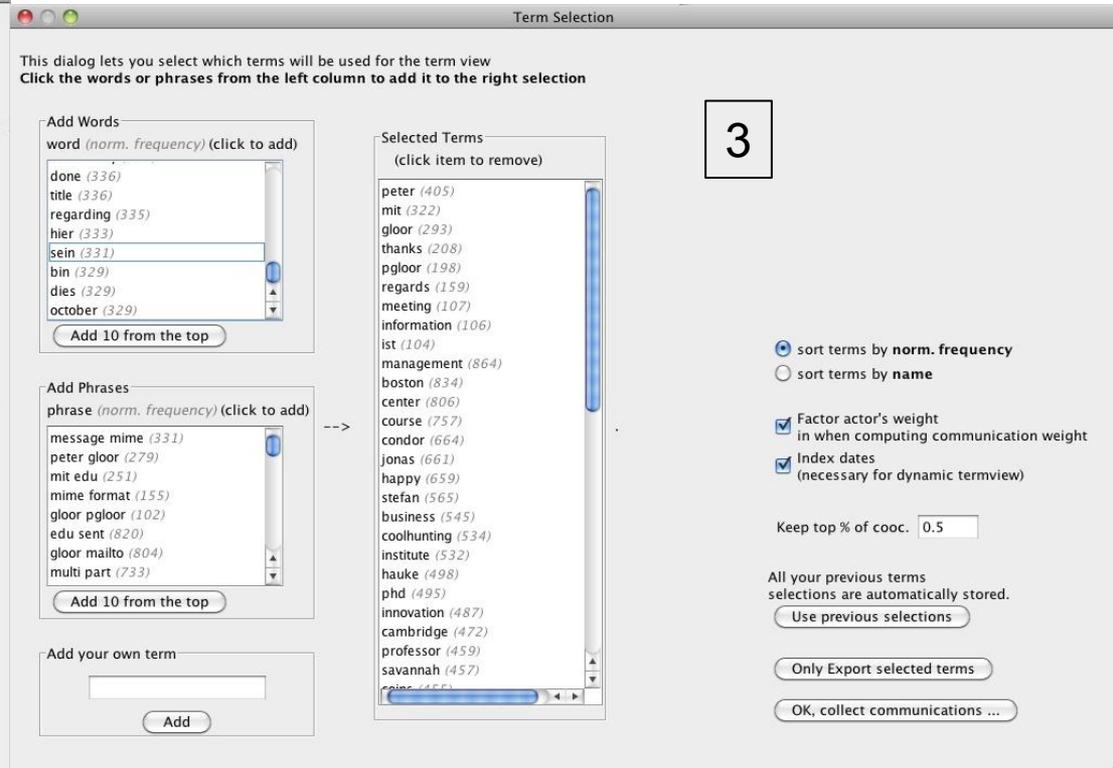
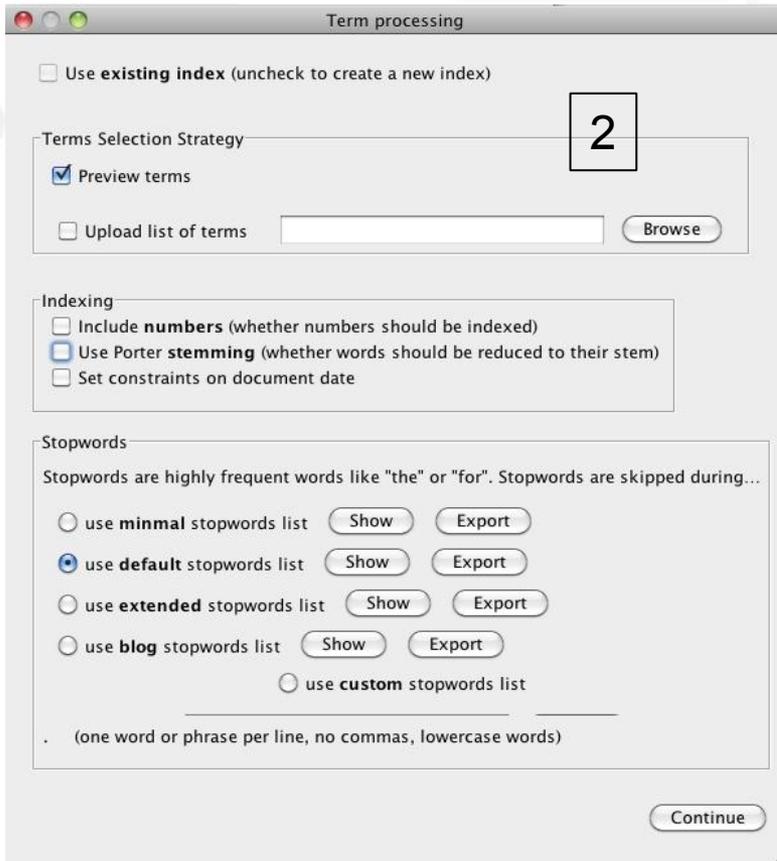
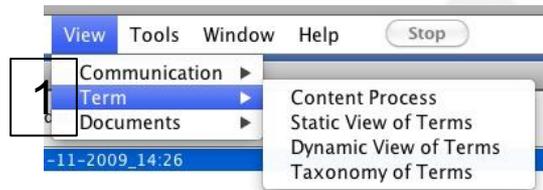
Visualize E-Mail Data (3)



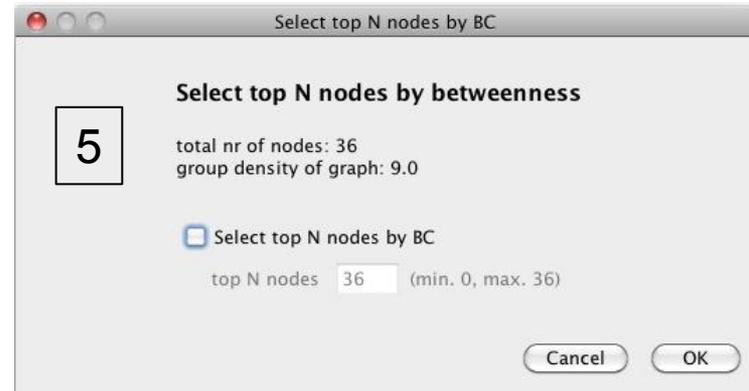
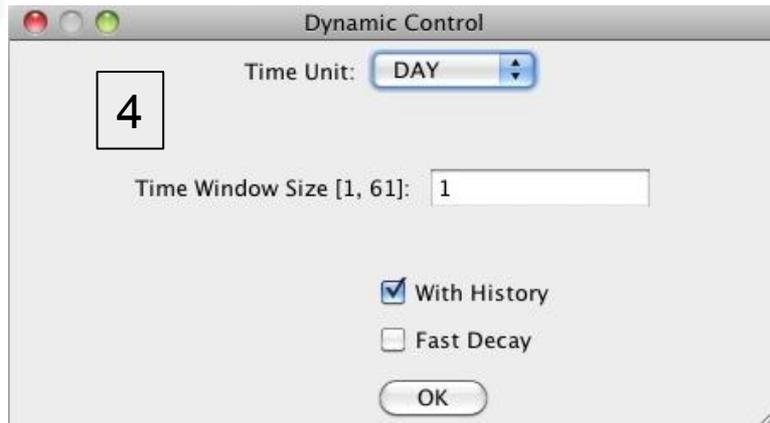
Dynamic View of Communication



Visualize E-Mail Contents



Visualize E-Mail Contents (2)



Contents

- Getting Started
- Collecting Web Content
- OneDegreeCollector
- Building your own Startlists
- Collecting your E-Mail
- **Collecting Facebook Data**
- **Collecting Wikipedia Data**
- **Collecting CoolPeople**
- Coolhunting Blueprint

MIT OpenCourseWare
<http://ocw.mit.edu>

15.599 Workshop in IT: Collaborative Innovation Networks
Fall 2011

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.