Evaluating Similarity Measures for Emergent Semantics of Social Tagging

Ben Markines, Ciro Cattuto,
Fil Menczer, Dominik Benz,
Andreas Hotho, Gerd Stumme

INDIANA UNIVERSITY
SCHOOL OF INFORMATICS
Bloomington

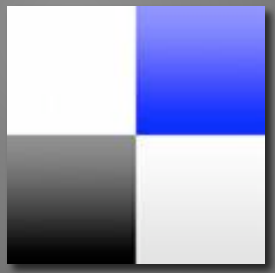
ISI Foundation

UNIKASSEL
VERSITÄT

Social Applications

BibSonomy

myspace.com
a place for friends.

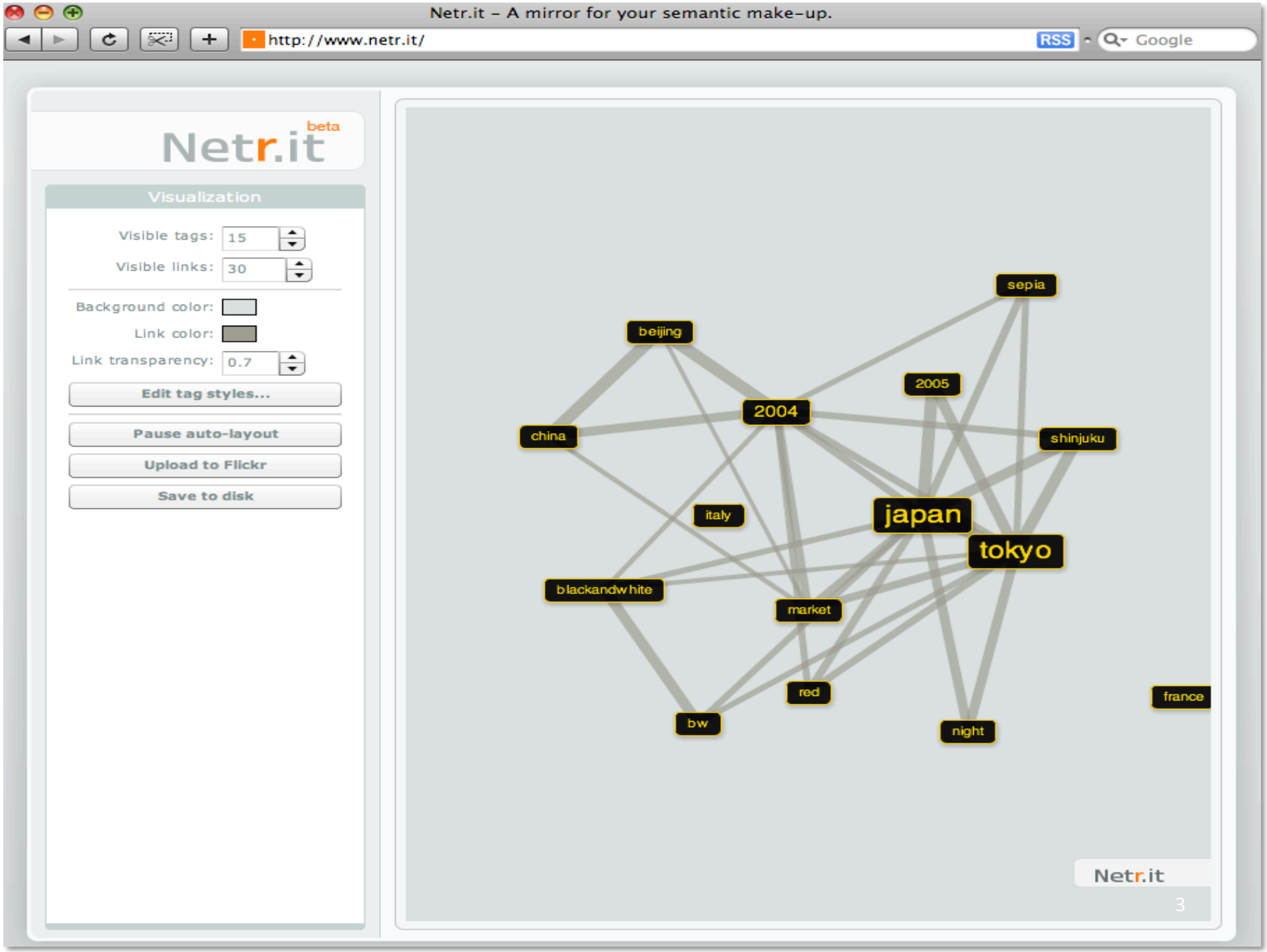


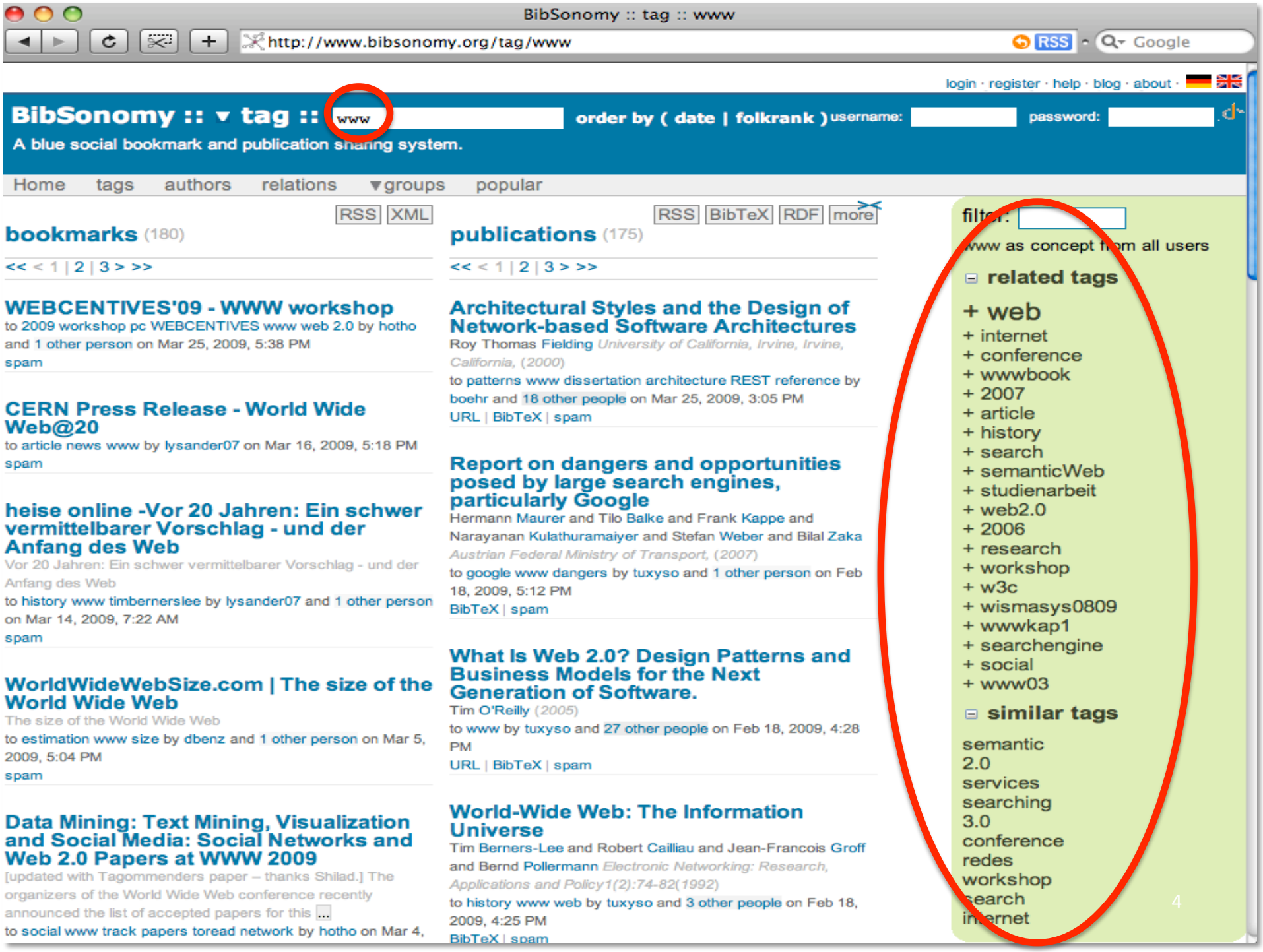
YouTube
Broadcast Yourself

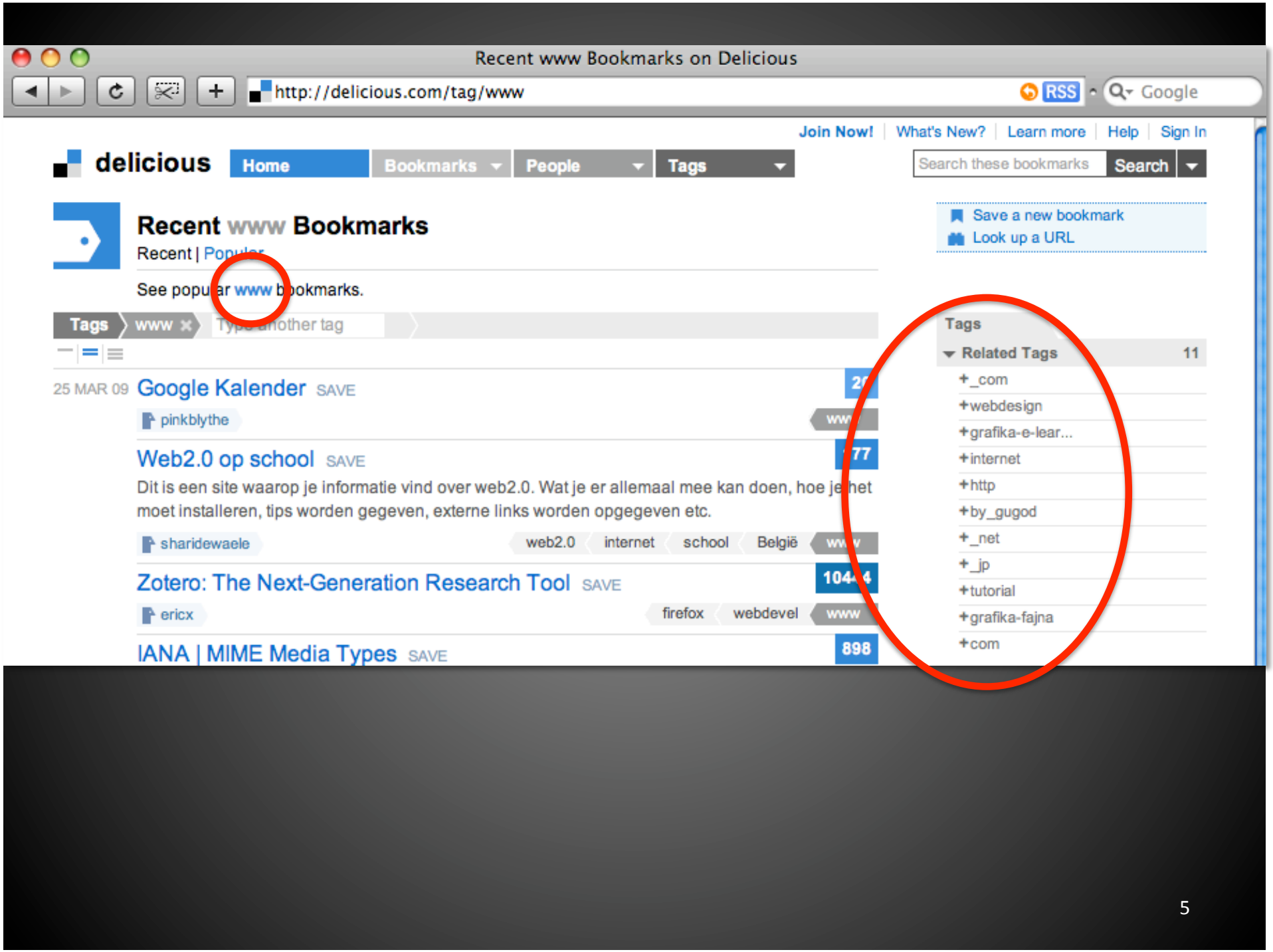
LinkedIn

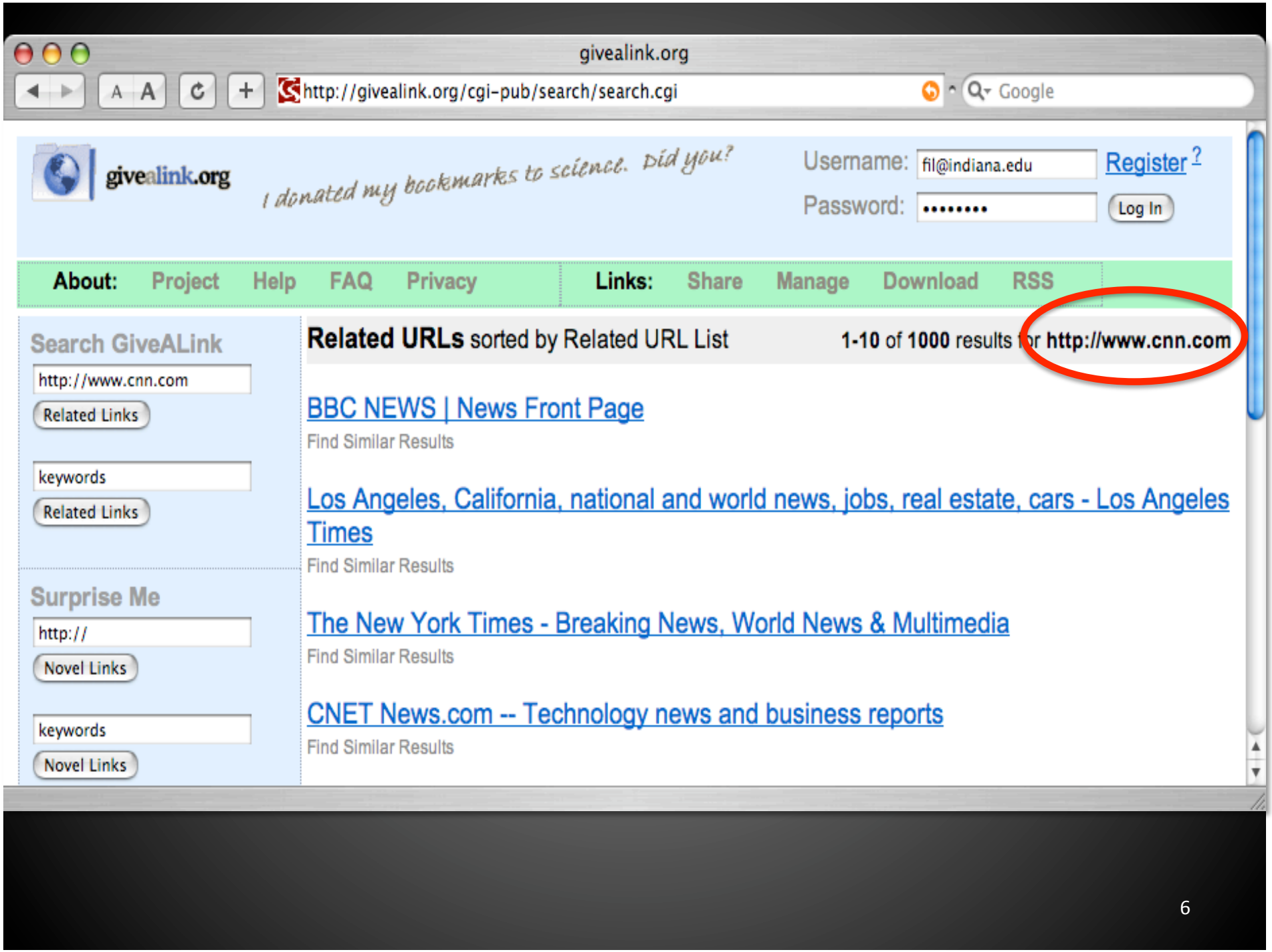
facebook

flickr


Netr.it – A mirror for your semantic make-up.
http://www.netr.it/
RSS
Google
Netr.it
beta
Visualization
Visible tags: 15
Visible links: 30
Background color:
Link color:
Link transparency: 0.7
Edit tag styles...
Pause auto-layout
Upload to Flickr
Save to disk
sepia
beijing
2005
2004
china
shinjuku
italy
japan
tokyo
blackandwhite
market
red
france
bw
night
Netr.it


BibSonomy :: tag :: www
http://www.bibsonomy.org/tag/www
RSS
Google
login · register · help · blog · about ·
BibSonomy :: tag :: www
order by (date | folkrank) username: password:
A blue social bookmark and publication sharing system.
Home tags authors relations groups popular
RSS XML
bookmarks (180)
<< < 1 | 2 | 3 > >>
WEBCENTIVES'09 - WWW workshop
to 2009 workshop pc WEBCENTIVES www web 2.0 by hotho and 1 other person on Mar 25, 2009, 5:38 PM
spam
CERN Press Release - World Wide Web@20
to article news www by lysander07 on Mar 16, 2009, 5:18 PM
spam
heise online -Vor 20 Jahren: Ein schwer vermittelbarer Vorschlag - und der Anfang des Web
Vor 20 Jahren: Ein schwer vermittelbarer Vorschlag - und der Anfang des Web
to history www timbernerslee by lysander07 and 1 other person on Mar 14, 2009, 7:22 AM
spam
WorldWideWebSize.com | The size of the World Wide Web
The size of the World Wide Web
to estimation www size by dbenz and 1 other person on Mar 5, 2009, 5:04 PM
spam
Data Mining: Text Mining, Visualization and Social Media: Social Networks and Web 2.0 Papers at WWW 2009
[updated with Tagommenders paper -- thanks Shilad.] The organizers of the World Wide Web conference recently announced the list of accepted papers for this ...
to social www track papers toread network by hotho on Mar 4,
RSS BibTeX RDF more
publications (175)
<< < 1 | 2 | 3 > >>
Architectural Styles and the Design of Network-based Software Architectures
Roy Thomas Fielding University of California, Irvine, Irvine, California, (2000)
to patterns www dissertation architecture REST reference by boehr and 18 other people on Mar 25, 2009, 3:05 PM
URL | BibTeX | spam
Report on dangers and opportunities posed by large search engines, particularly Google
Hermann Maurer and Tilo Balke and Frank Kappe and Narayanan Kulathuramaiyer and Stefan Weber and Bilal Zaka Austrian Federal Ministry of Transport, (2007)
to google www dangers by tuxyso and 1 other person on Feb 18, 2009, 5:12 PM
BibTeX | spam
What Is Web 2.0? Design Patterns and Business Models for the Next Generation of Software.
Tim O'Reilly (2005)
to www by tuxyso and 27 other people on Feb 18, 2009, 4:28 PM
URL | BibTeX | spam
World-Wide Web: The Information Universe
Tim Berners-Lee and Robert Cailliau and Jean-Francois Groff and Bernd Pollermann Electronic Networking: Research, Applications and Policy1(2):74-82(1992)
to history www web by tuxyso and 3 other people on Feb 18, 2009, 4:25 PM
BibTeX | spam
filter:
www as concept from all users
related tags
+ web
+ internet
+ conference
+ wwwbook
+ 2007
+ article
+ history
+ search
+ semanticWeb
+ studienarbeit
+ web2.0
+ 2006
+ research
+ workshop
+ w3c
+ wismasys0809
+ wwwkap1
+ searchengine
+ social
+ www03
similar tags
semantic
2.0
services
searching
3.0
conference
redes
workshop
search
internet

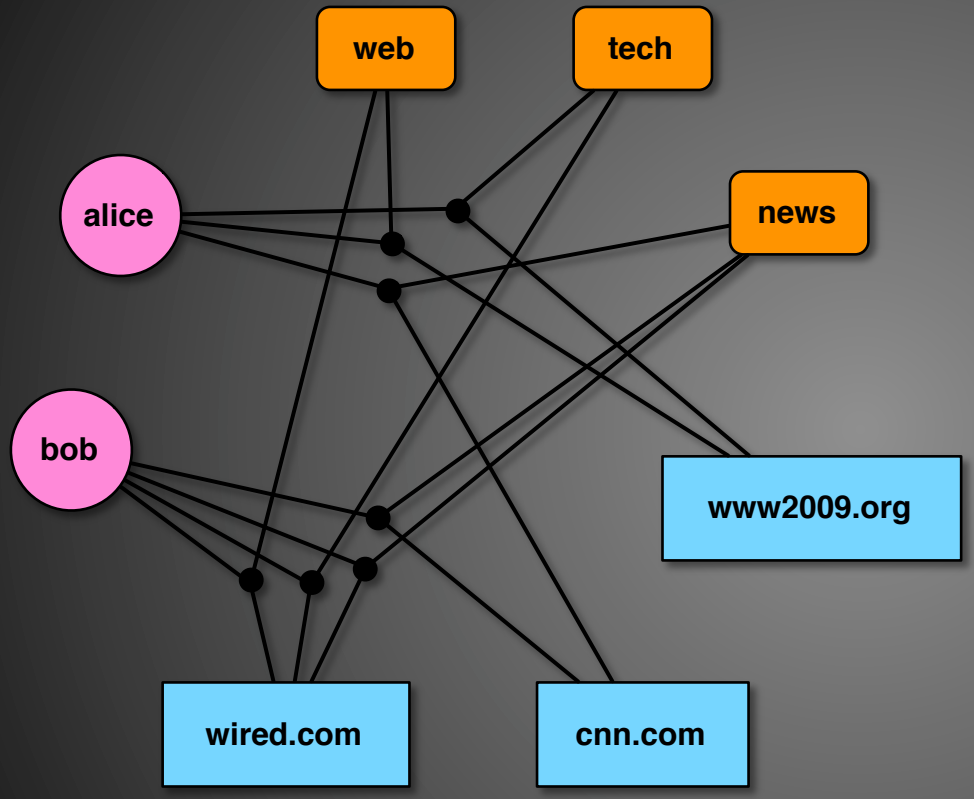

Recent www Bookmarks on Delicious
http://delicious.com/tag/www
RSS
Google
Join Now! | What's New? | Learn more | Help | Sign In
delicious
Home
Bookmarks
People
Tags
Search these bookmarks
Search
Recent www Bookmarks
Recent | Popular
See popular www bookmarks.
Save a new bookmark
Look up a URL
Tags
www
Type another tag
25 MAR 09
Google Kalender SAVE
pinkblythe
Web2.0 op school SAVE
Dit is een site waarop je informatie vind over web2.0. Wat je er allemaal mee kan doen, hoe je het moet installeren, tips worden gegeven, externe links worden opgegeven etc.
sharidewaele
web2.0
internet
school
België
Zotero: The Next-Generation Research Tool SAVE
ericx
firefox
webdevel
IANA | MIME Media Types SAVE
898
Tags
Related Tags 11
+_com
+webdesign
+grafika-e-lear...
+internet
+http
+by_gugod
+_net
+_jp
+tutorial
+grafika-fajna
+com


givealink.org
http://givealink.org/cgi-pub/search/search.cgi
Google
givealink.org
I donated my bookmarks to science. Did you?
Username: fil@indiana.edu
Register ?
Password: ••••••••
Log In
About: Project Help FAQ Privacy
Links: Share Manage Download RSS
Search GiveALink
http://www.cnn.com
Related Links
keywords
Related Links
Surprise Me
http://
Novel Links
keywords
Novel Links
Related URLs sorted by Related URL List
1-10 of 1000 results for http://www.cnn.com
BBC NEWS | News Front Page
Find Similar Results
Los Angeles, California, national and world news, jobs, real estate, cars - Los Angeles Times
Find Similar Results
The New York Times - Breaking News, World News & Multimedia
Find Similar Results
CNET News.com -- Technology news and business reports
Find Similar Results

Goals

- **Tag-tag**, **resource-resource**, user-user similarity
- Capture relationships
  - Effectively and Efficiently
  - Shannon information of annotations

Folksonomy Model



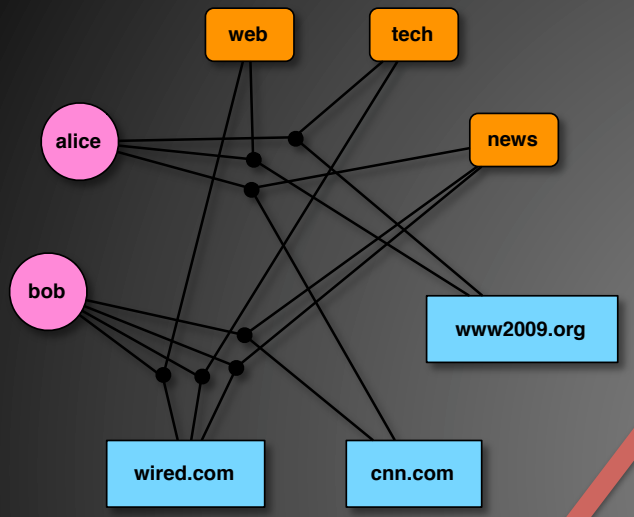
- hyper-graph
- complex
- user-driven
- large-scale
- many-projections
- literature

$$F = (U, T, R, Y), Y \subseteq U \times T \times R \text{ (the triples)}$$

Agenda

- **Design**
  - **Aggregation**
  - **Similarity Measures**
- Evaluation

Aggregation Methods



alice

	news	web	tech
cnn.com	1	0	0
www2009.org	0	1	1

bob

	news	web	tech
cnn.com	1	0	0
wired.com	1	1	1

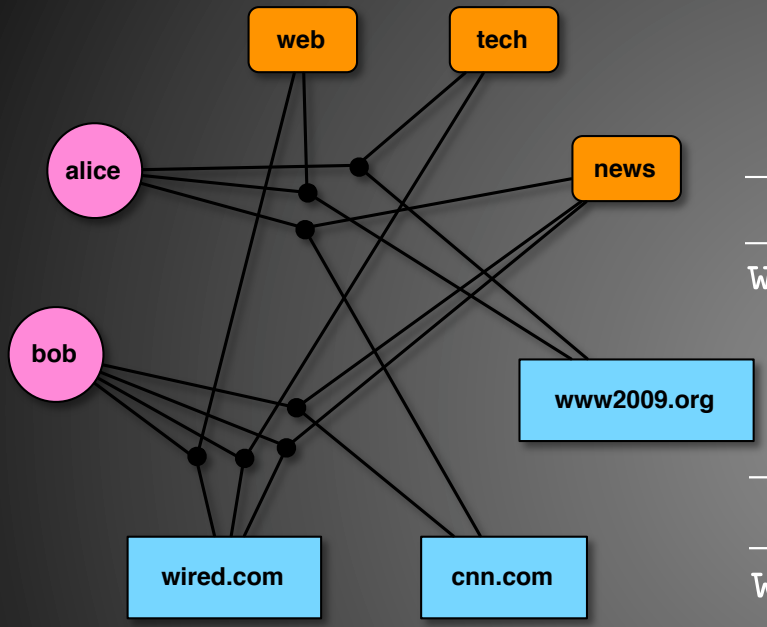
Projection

	news	web	tech
cnn.com	1	0	0
www2009.org	0	1	1
wired.com	1	1	1

Distributional

	news	web	tech
cnn.com	2	0	0
www2009.org	0	1	1
wired.com	1	1	1

Aggregation Methods: Incremental



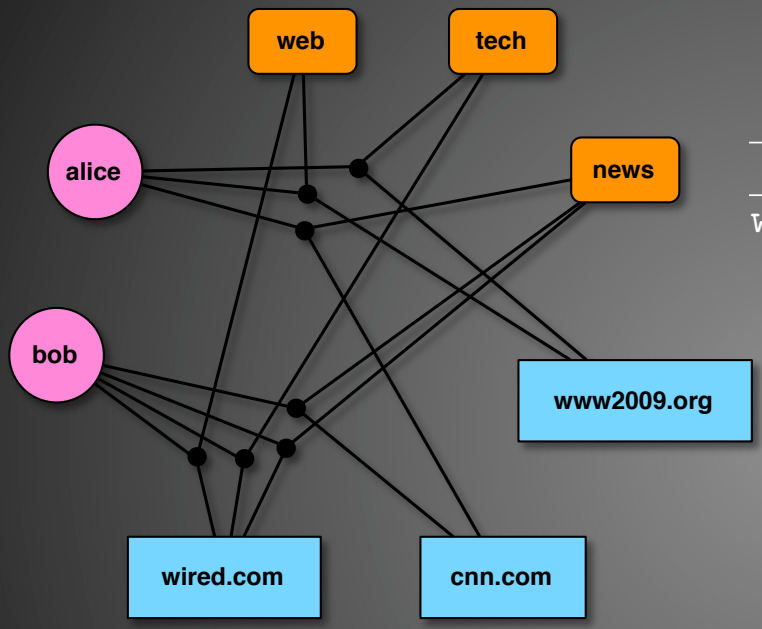
Macro

	news	web	tech	
cnn.com	1	0	0	alice
www2009.org	0	1	1	

	news	web	tech	
cnn.com	1	0	0	bob
wired.com	1	1	1	

$$\sigma(x, y) = \sum_u \sigma_u(x, y)$$

Aggregation Methods: Incremental (2)



Collaborative

	news	web	tech	alice
cnn.com	1	0	0	1
www2009.org	0	1	1	1

alice

	news	web	tech	bob
cnn.com	1	0	0	1
wired.com	1	1	1	1

bob

$$\sigma(x, y) = \sum_u \sigma_u(x, y)$$

Similarity Measures

- **Jaccard**
- Matching
- Overlap
- Dice

$$\sigma(x_1, x_2) = \frac{|X_1 \cap X_2|}{|X_1 \cup X_2|}$$

attribute vector

object

attribute

projection

$$\sigma(x_1, x_2) = \frac{\sum_{y \in X_1 \cap X_2} \log p(y)}{\sum_{y \in X_1 \cup X_2} \log p(y)}$$

distributional

attribute probability

$$\sigma_u(x_1, x_2) = \frac{\sum_{y \in X_1^u \cap X_2^u} \log p(y|u)}{\sum_{y \in X_1^u \cup X_2^u} \log p(y|u)}$$

user

macro/collaborative

$$p(y|u) = \frac{N(u, y)}{N(u) + \delta}$$

More Similarity Measures

- Cosine

$$\sigma(x_1, x_2) = \frac{X_1 \bullet X_2}{||X_1||\ ||X_2||}$$

- Mutual Information

$$\sigma(x_1, x_2) = \sum_{y_1 \in X_1} \sum_{y_2 \in X_2} p(y_1, y_2) \log \frac{p(y_1, y_2)}{p(y_1)p(y_2)}$$

attribute joint probability

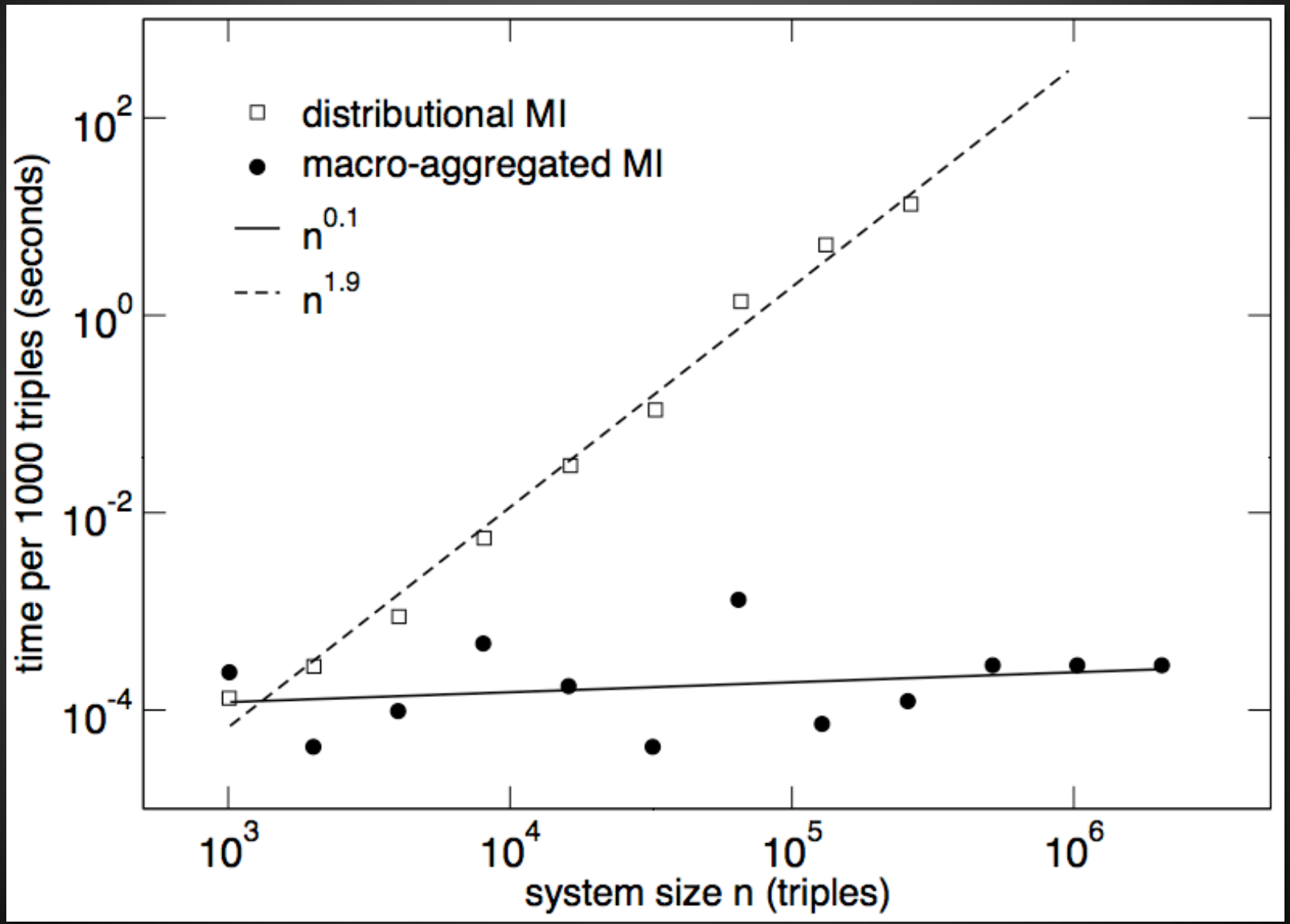
- Maximum Information Path

$$\sigma(x_1, x_2) = \frac{2 \times \log(\min_{y \in X_1 \cap X_2}[p(y)])}{\log(\min_{y \in X_1}[p(y)]) + \log(\min_{y \in X_2}[p(y)])}$$

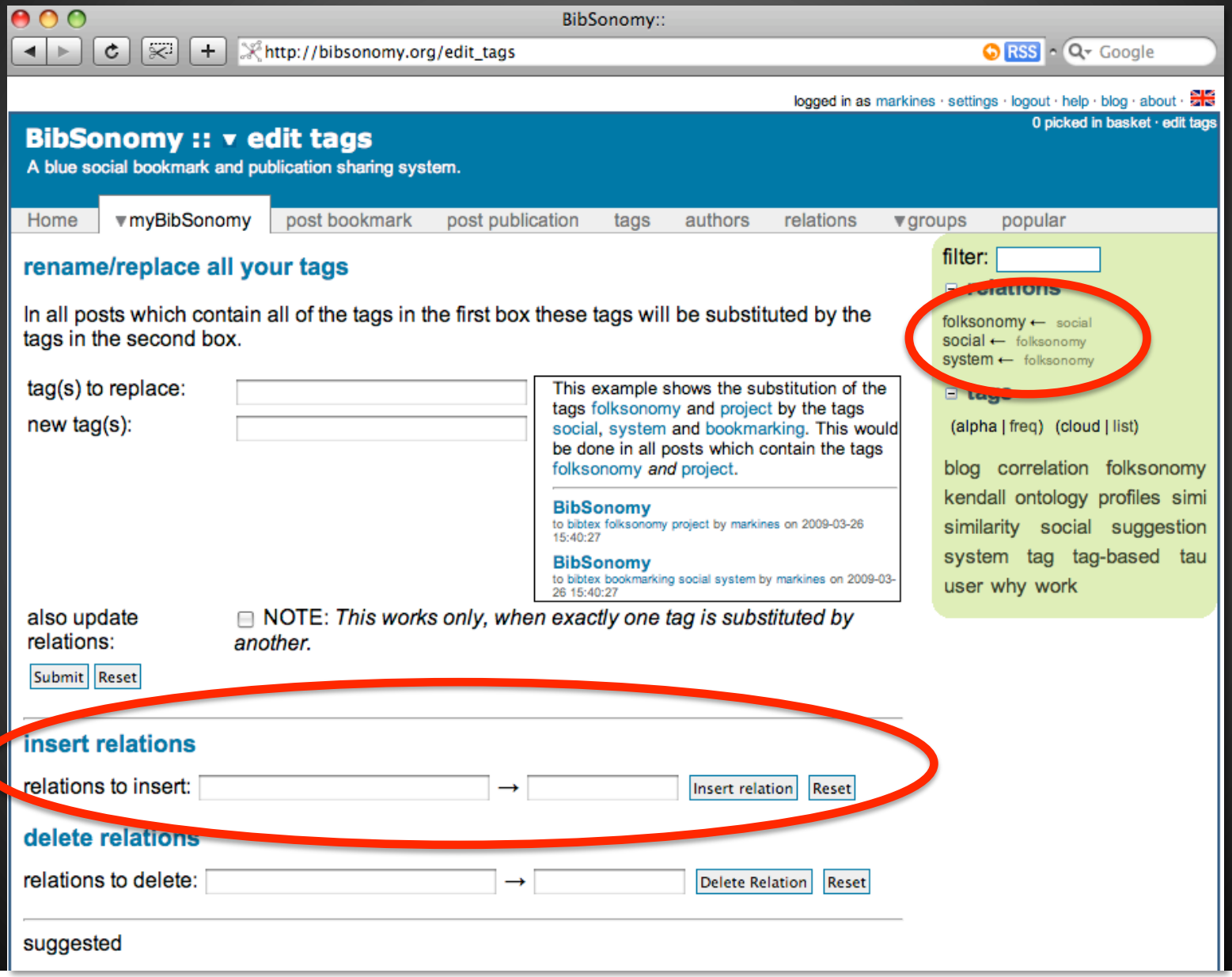
Agenda

- Design
- **Evaluation**
  - **Efficiency**
  - **Predicting tag relations**
  - **Semantic Grounding**

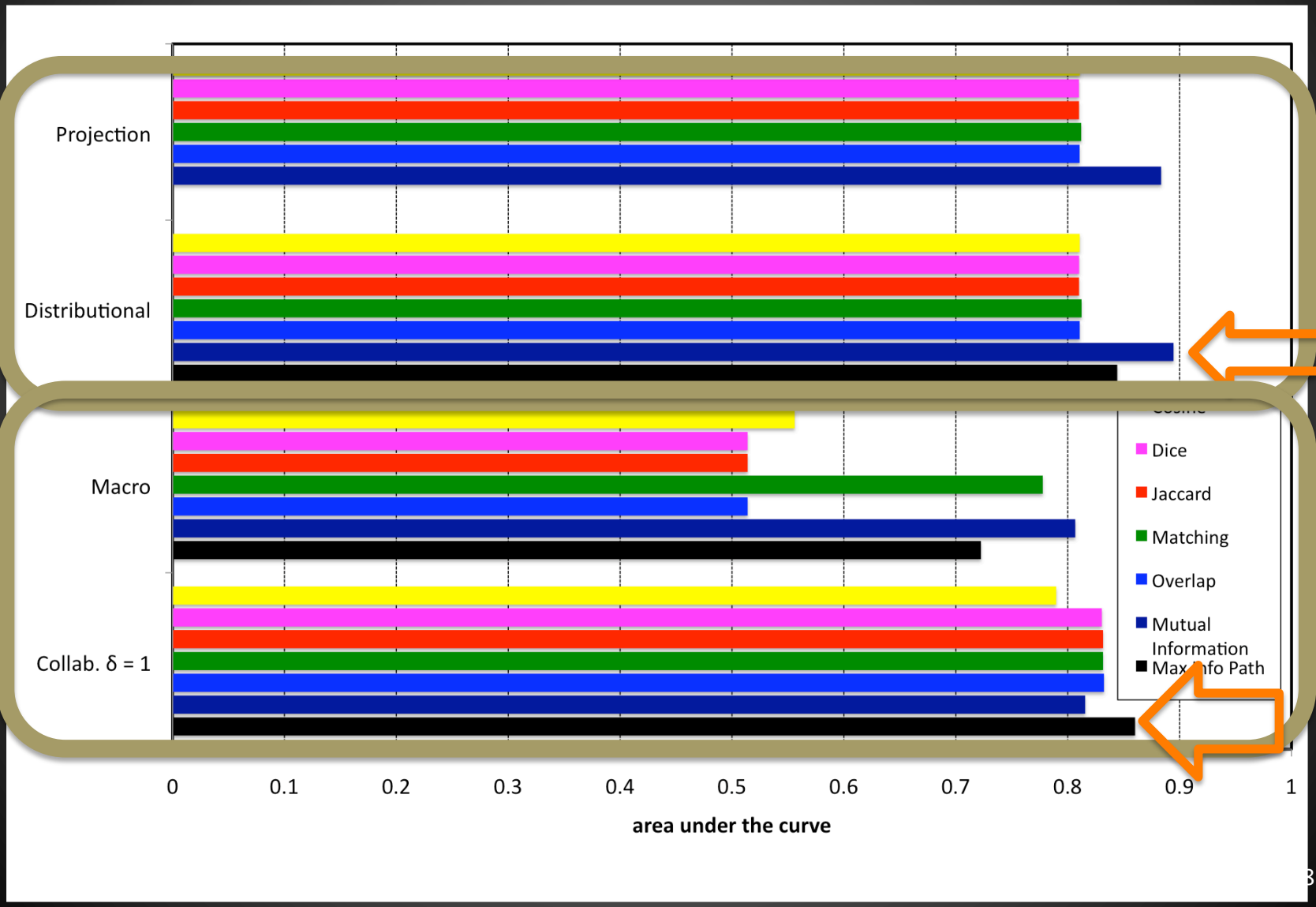
Update Time


distributional MI
macro-aggregated MI
$n^{0.1}$
$n^{1.9}$
10^2
10^0
10^{-2}
10^{-4}
time per 1000 triples (seconds)
10^3
10^4
10^5
10^6
system size n (triples)

Predicting User-defined Tag Relations



Predicting User-defined Tag Relations Area Under ROC Curve



ROC limitations

- Data is sparse: 2,000 tags
  - 142 user tag relations
- Similarity values are broadly distributed
- Tag relations rely on hierarchical relationships
- Only available with tags (not resources)

Semantic Grounding

WordNet a lexical database for the English language

- 17,041 tags
  - Overlap between WordNet and Bibsonomy tags
- Limited to the top 2,000 resources
- Relationships established with Jiang-Conrath
  - user-validated

Jiang ROCLING 1997

dmoz open directory project

- 3,323 resources
  - Overlap between ODP and Bibsonomy resources
- Relationships established with Maguitman's graph based similarity
  - user-validated

Maguitman WWW 2005

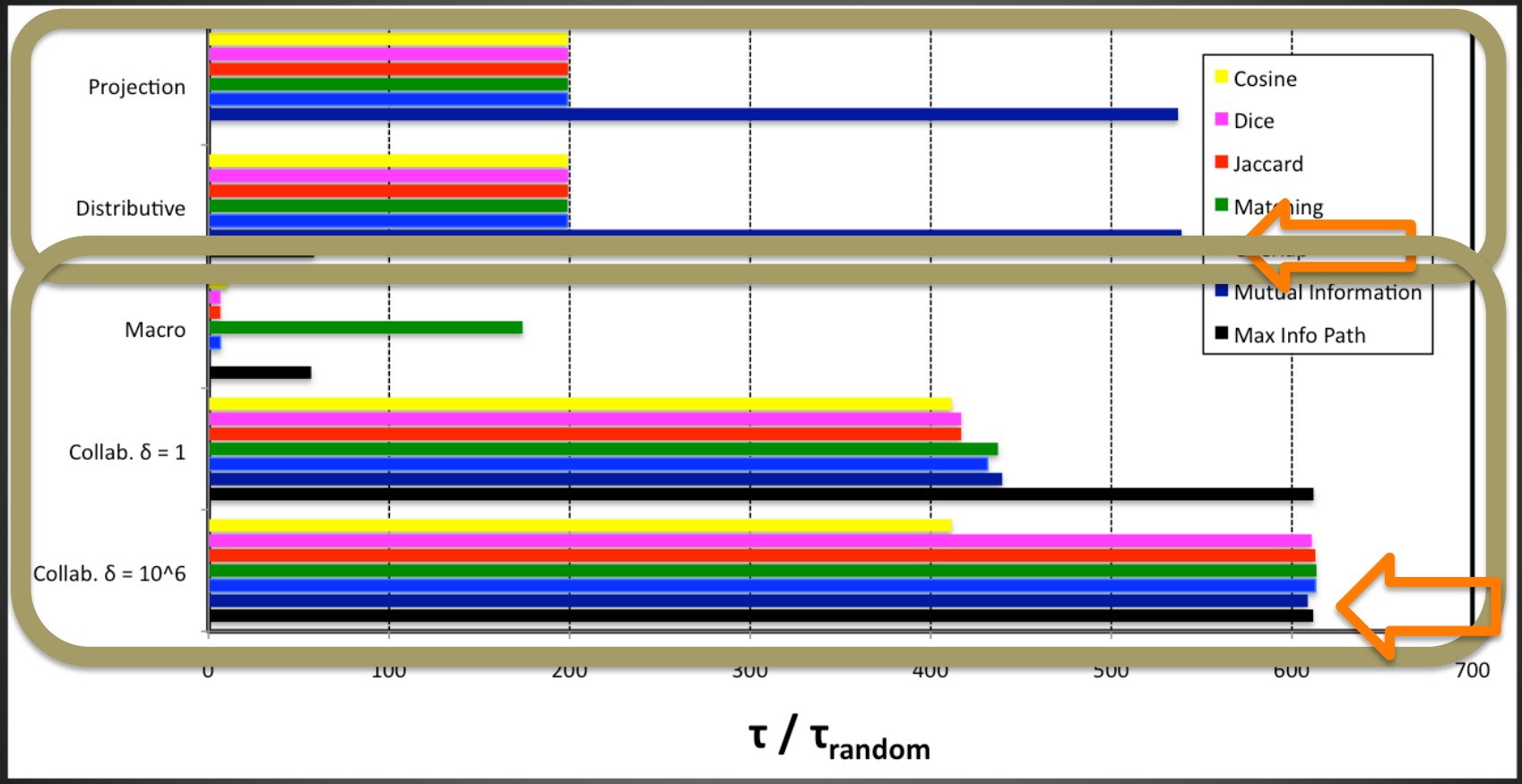
Kendall's τ

Rank	Reference	Measure A	Measure B
1	`tech-web`	`news-tech`	`news-web`
2	`news-web`	`tech-web`	`tech-web`
3	`news-tech`	`news-web`	`news-tech`
τ	1	1/3	2/3

$$\tau = \frac{|\text{agreed ranked pairs}|}{|\text{total number of ranked pairs}|}$$

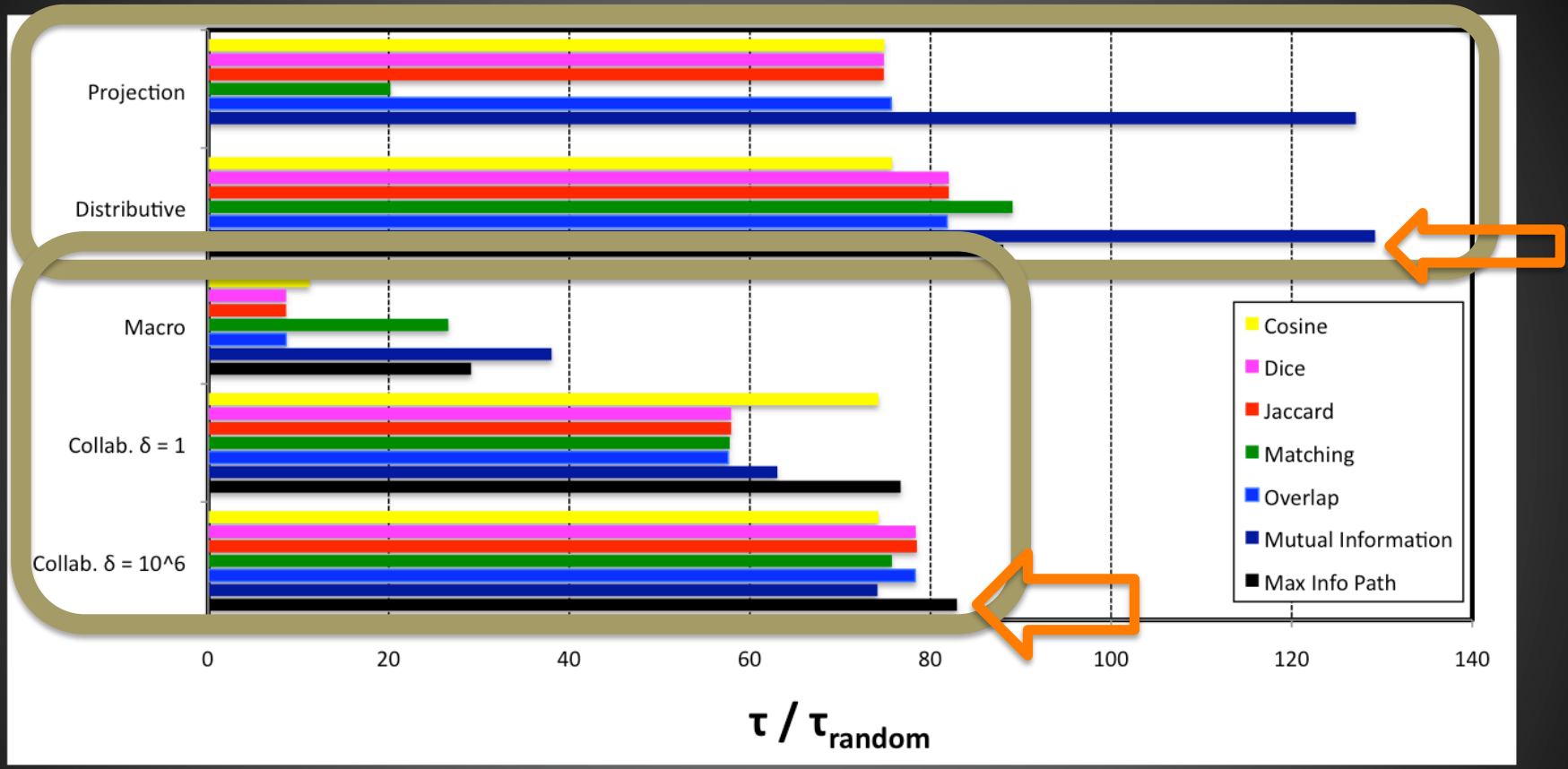
Kendall Biometrika 1938

Tag Similarity



random $\tau = 10^{-4}$

Resource Similarity



random $\tau = 8 \times 10^{-5}$

Related Work

- User, tag and resource similarity
  - Mika 2005, Cattuto et al. 2008, Wu et al. 2006, Diederich and Iofciu 2006
- Ranking
  - Hotho et al. 2006
- Organization
  - Begelman et al. 2006, Sigurbjörnsson and van Zwol 2008, Heymann and Garcia-Molina 2006
- Link Prediction
  - Liben-Nowell and Kleinberg 2003
- Recommendation
  - Mishne 2006, Brooks and Montanez 2006, Sarwar et al. 2001

Conclusion

- Similarity framework
  - Folksonomy-based tag/resource similarity measures
  - Aggregation methods
- Evaluation
  - Efficiency/performance tradeoffs
  - Direct vs. semantic grounding
  - Distributional Mutual Information performs well, but is inefficient
  - Collaborative aggregation is both efficient and effective, especially Maximum Information Path
- Techniques presented here can immediately support Social Web applications

Thank You!

- Similarity framework
  - Folksonomy-based tag/resource similarity measures
  - Aggregation methods
- Evaluation
  - Efficiency/performance tradeoffs
  - Direct vs. semantic grounding
  - Distributional Mutual Information performs well, but is inefficient
  - Collaborative aggregation is both efficient and effective, especially Maximum Information Path
- Techniques presented here can immediately support Social Web applications

Ben Markines

Ciro Cattuto

Fil Menczer

Dominik Benz

Andreas Hotho

Gerd Stumme