Ranking and Community detection in undirected networks

Andrea Baldassarri, Ciro Cattuto
Vittorio Loreto, Vito D.P. Servedio

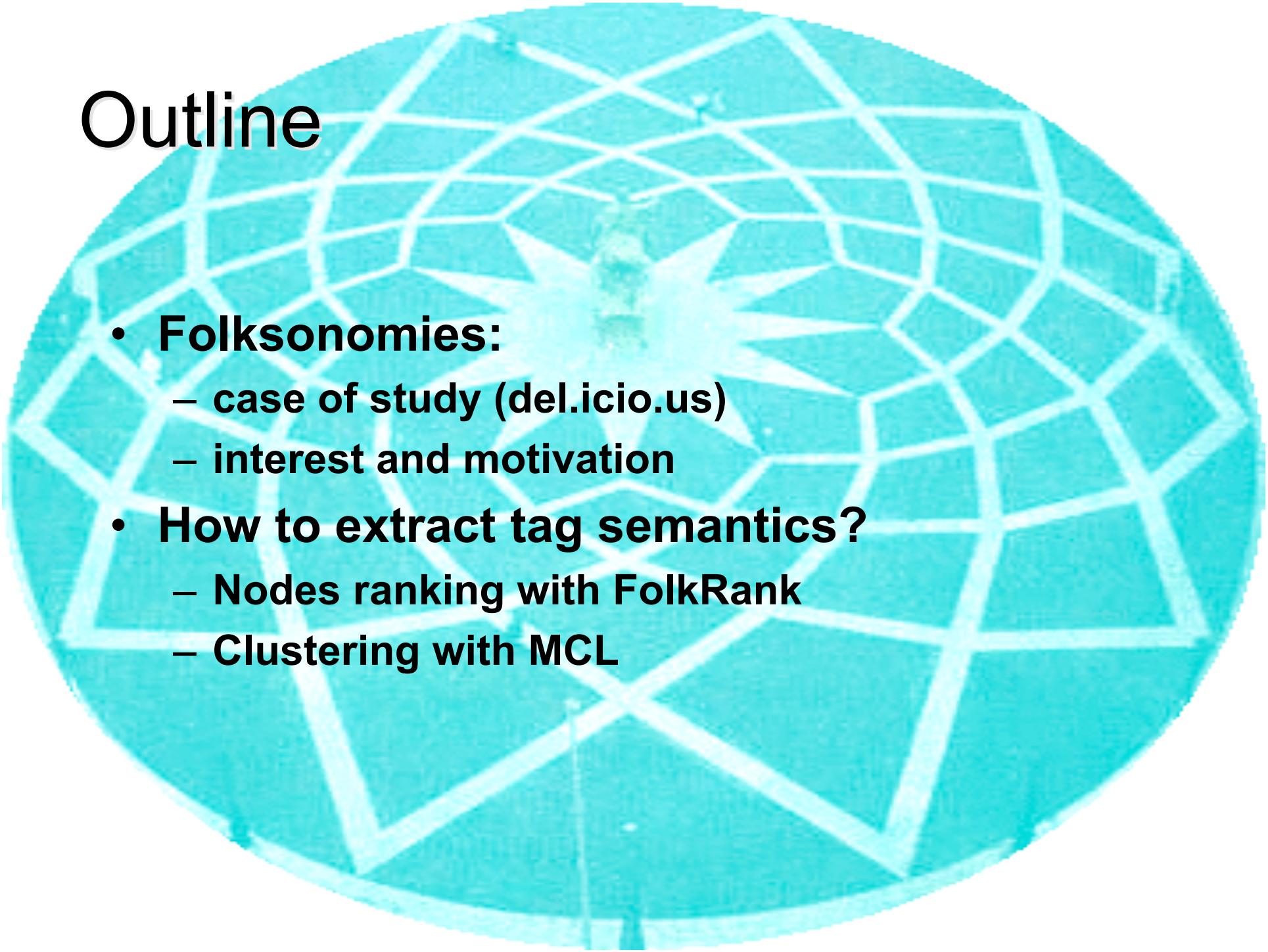
La Sapienza
Università degli Studi di Roma

ENDOWED CHAIR OF THE HERTIE FOUNDATION
Knowledge and Data Engineering
DEPARTMENT OF MATHEMATICS & COMPUTER SCIENCE

Miranda Grahl, Andreas Hotho
Christoph Schmitz, Gerd Stumme

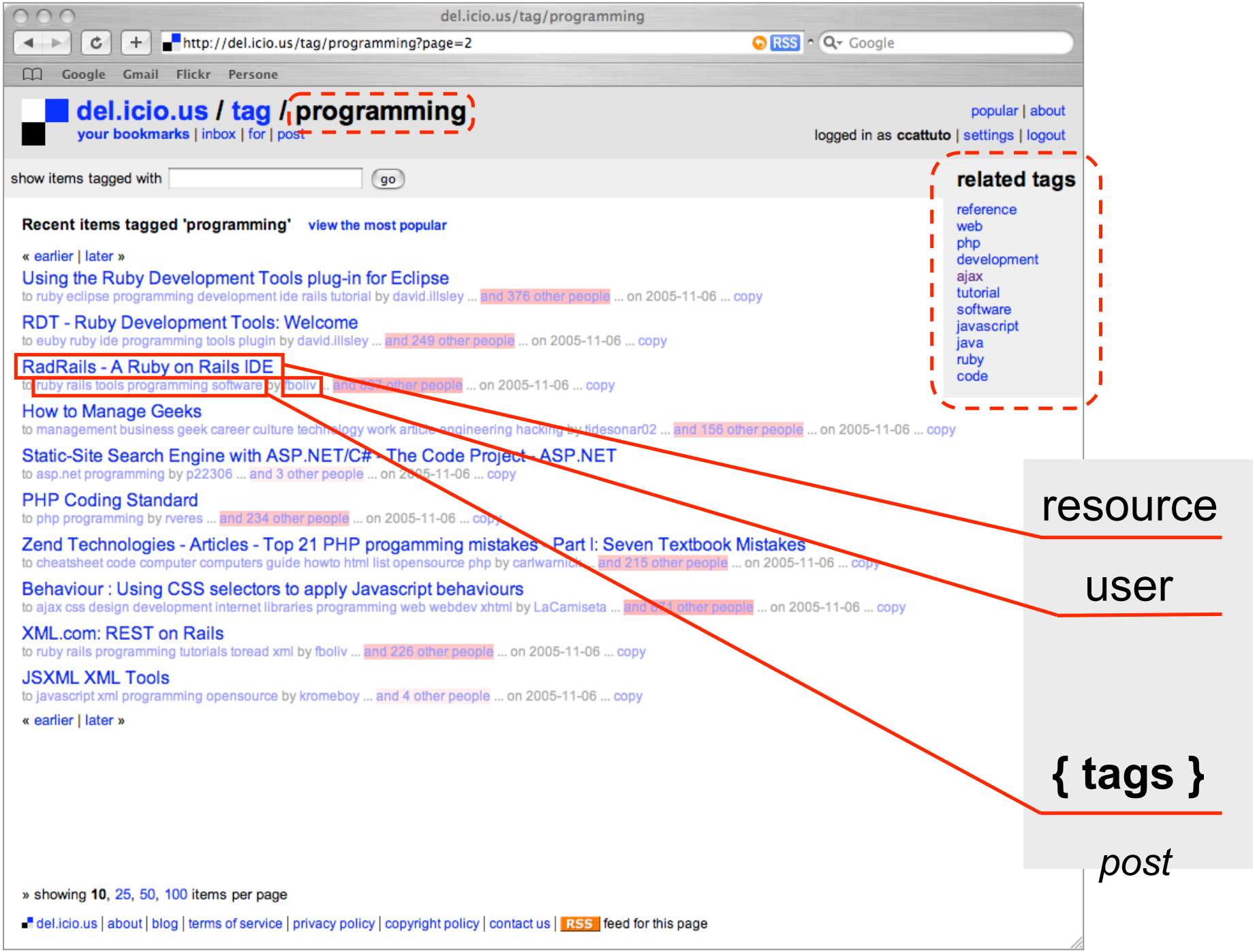
Tagora

http://www.tagora-project.eu

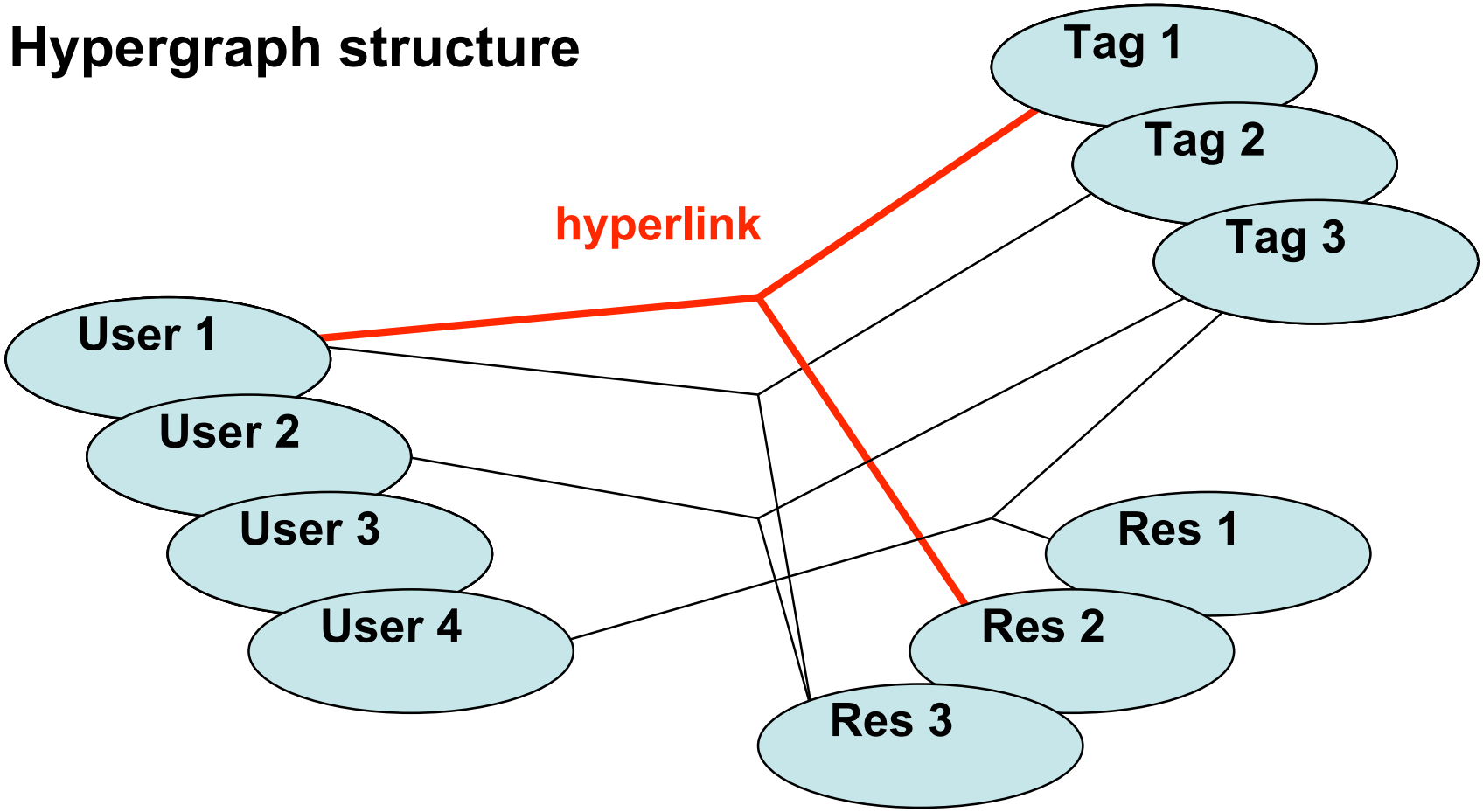


Outline

- **Folksonomies:**
  - **case of study (del.icio.us)**
  - **interest and motivation**
- **How to extract tag semantics?**
  - **Nodes ranking with FolkRank**
  - **Clustering with MCL**


del.icio.us/tag/programming
http://del.icio.us/tag/programming?page=2
RSS
Google
Gmail
Flickr
Persone
del.icio.us / tag / programming
your bookmarks | inbox | for | post
popular | about
logged in as ccattuto | settings | logout
show items tagged with
go
related tags
reference
web
php
development
ajax
tutorial
software
javascript
java
ruby
code
Recent items tagged 'programming'
view the most popular
« earlier | later »
Using the Ruby Development Tools plug-in for Eclipse
to ruby eclipse programming development ide rails tutorial by david.illsley ... and 376 other people ... on 2005-11-06 ... copy
RDT - Ruby Development Tools: Welcome
to euby ruby ide programming tools plugin by david.illsley ... and 249 other people ... on 2005-11-06 ... copy
RadRails - A Ruby on Rails IDE
to ruby rails tools programming software by fboliv ... and 457 other people ... on 2005-11-06 ... copy
How to Manage Geeks
to management business geek career culture technology work article engineering hacking by tidesonar02 ... and 156 other people ... on 2005-11-06 ... copy
Static-Site Search Engine with ASP.NET/C# - The Code Project - ASP.NET
to asp.net programming by p22306 ... and 3 other people ... on 2005-11-06 ... copy
PHP Coding Standard
to php programming by rveres ... and 234 other people ... on 2005-11-06 ... copy
Zend Technologies - Articles - Top 21 PHP progamming mistakes - Part I: Seven Textbook Mistakes
to cheatsheet code computer computers guide howto html list opensource php by carlwarnick ... and 215 other people ... on 2005-11-06 ... copy
Behaviour : Using CSS selectors to apply Javascript behaviours
to ajax css design development internet libraries programming web webdev xhtml by LaCamiseta ... and 371 other people ... on 2005-11-06 ... copy
XML.com: REST on Rails
to ruby rails programming tutorials toread xml by fboliv ... and 226 other people ... on 2005-11-06 ... copy
JSXML XML Tools
to javascript xml programming opensource by kromeboy ... and 4 other people ... on 2005-11-06 ... copy
« earlier | later »
» showing 10, 25, 50, 100 items per page
del.icio.us | about | blog | terms of service | privacy policy | copyright policy | contact us | RSS feed for this page
resource
user
{ tags }
post

Folksonomy


Hypergraph structure
Tag 1
Tag 2
Tag 3
hyperlink
User 1
User 2
User 3
User 4
Res 1
Res 2
Res 3

Crawl statistics (KDE)

Crawled del.icio.us in July 2005

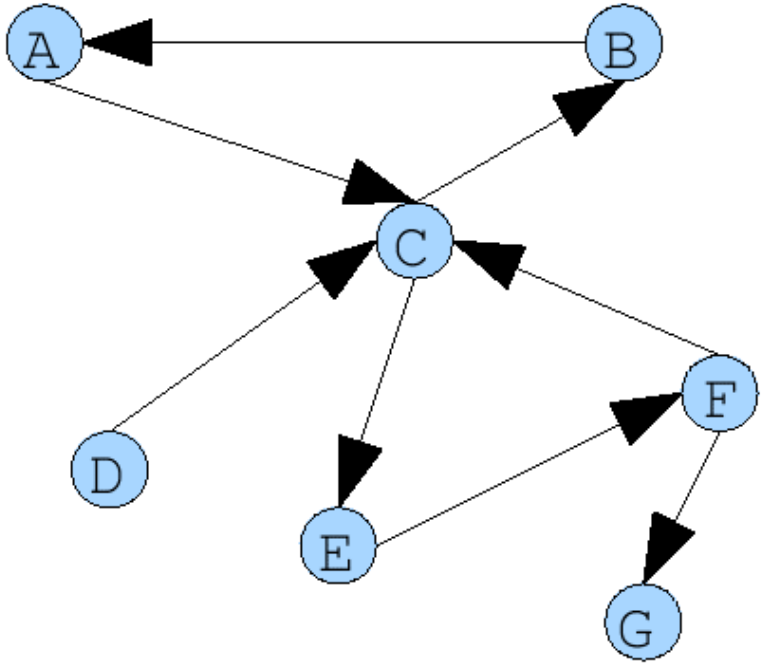
- 75,242 users
- 533,191 different tags
- 3,158,297 different resources
- 17,362,212 total tag assignments (TAS)

More recent crawl (TAGora, 2007):
> 600.000 users
> 12.000.000 resources
> 3.000.000 different tags

Ranking in directed graphs

PageRank:

I get high rank if pages with high rank point me



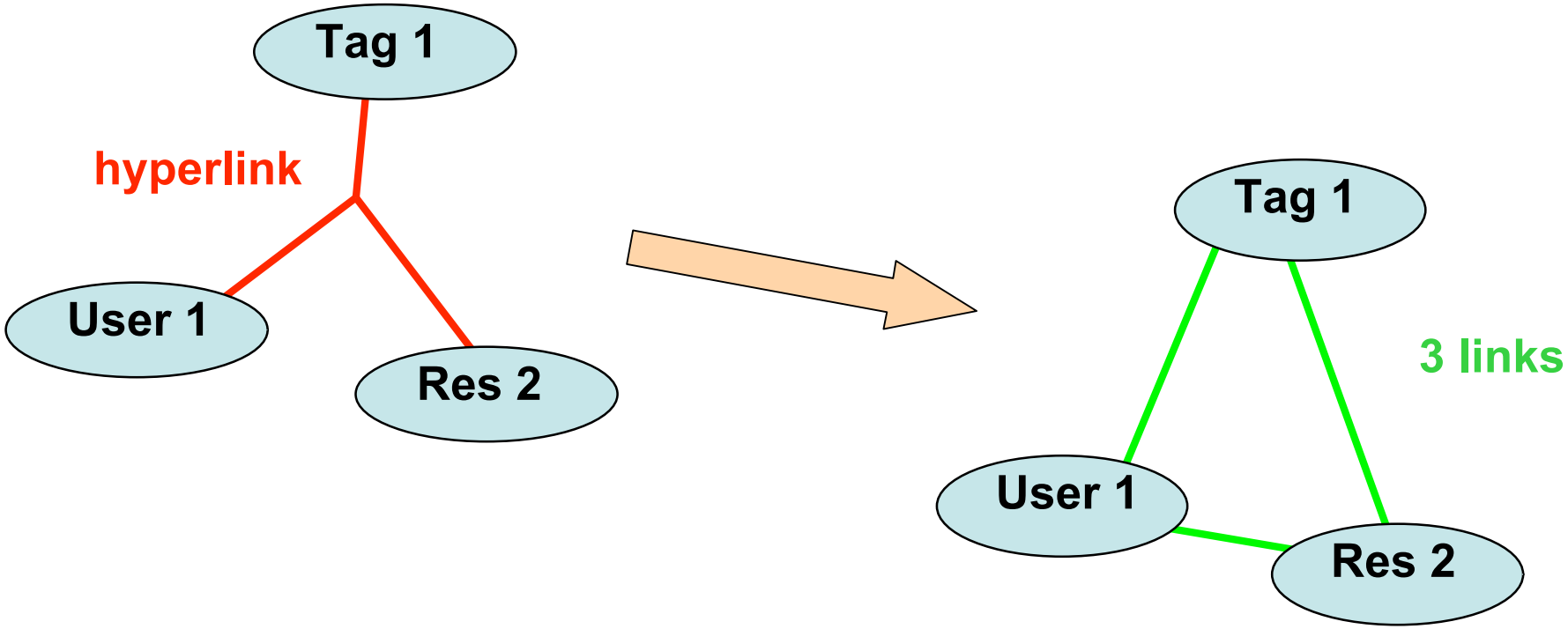
$$\vec{w} \leftarrow dA\vec{w} + (1-d)\vec{p}$$

d = 0.85, p uniform
A = graph normal matrix

Interpretation:
walker that jumps to a random location with probability 1-d; follows links with probability d.

Converting a Folksonomy into an undirected graph

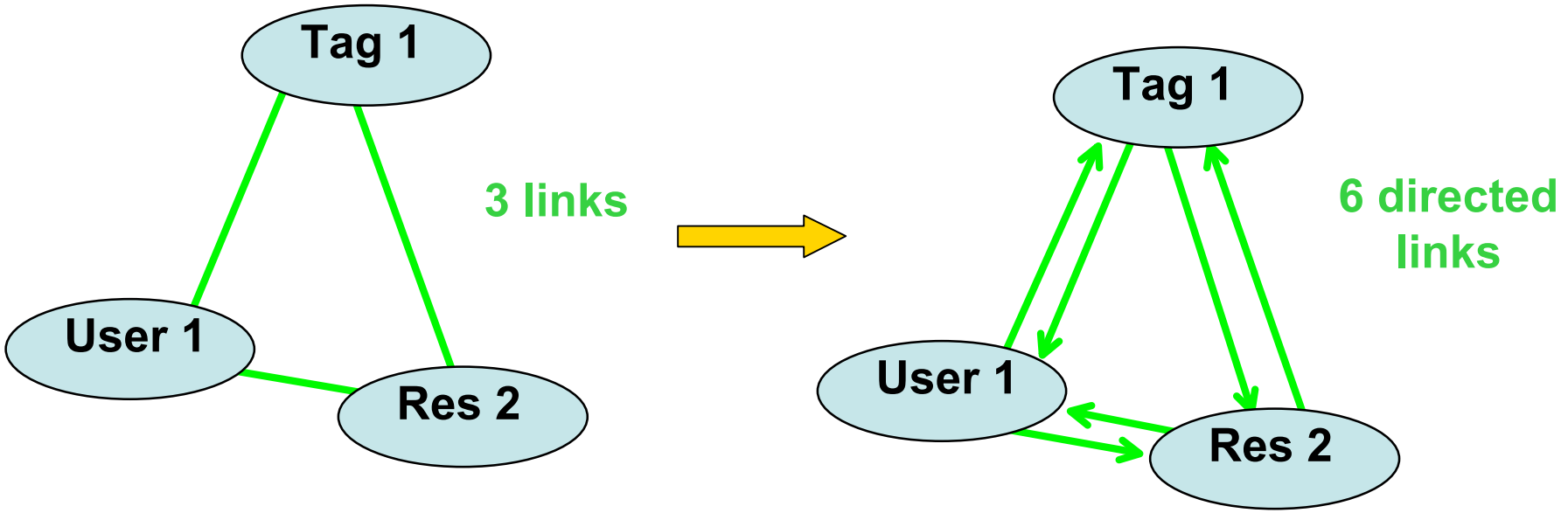
From the hypergraph to the undirected graph:



Substitute a hyperlink with 3 links

A Folksonomy adapted PageRank

1 undirected edge → 2 directed edges



PageRank of a node is proportional to its degree:
not a really useful semantic information

FolkRank:
Thematic Ranking in Folksonomies

Preferential PageRank: the walker prefers to jump to a given location

$$\vec{w} \leftarrow dA\vec{w} + (1-d)\vec{p}$$

d = 0.85, ***p*** ***preferential***
A = graph normal matrix

50% prob to jump to a given node

Frequency effects still present

FolkRank: differential ranking dumps frequency effects

$$W_{FolkRank} = W_{preferential} - W_{PageRank}$$

Comparing results of ranking algorithms

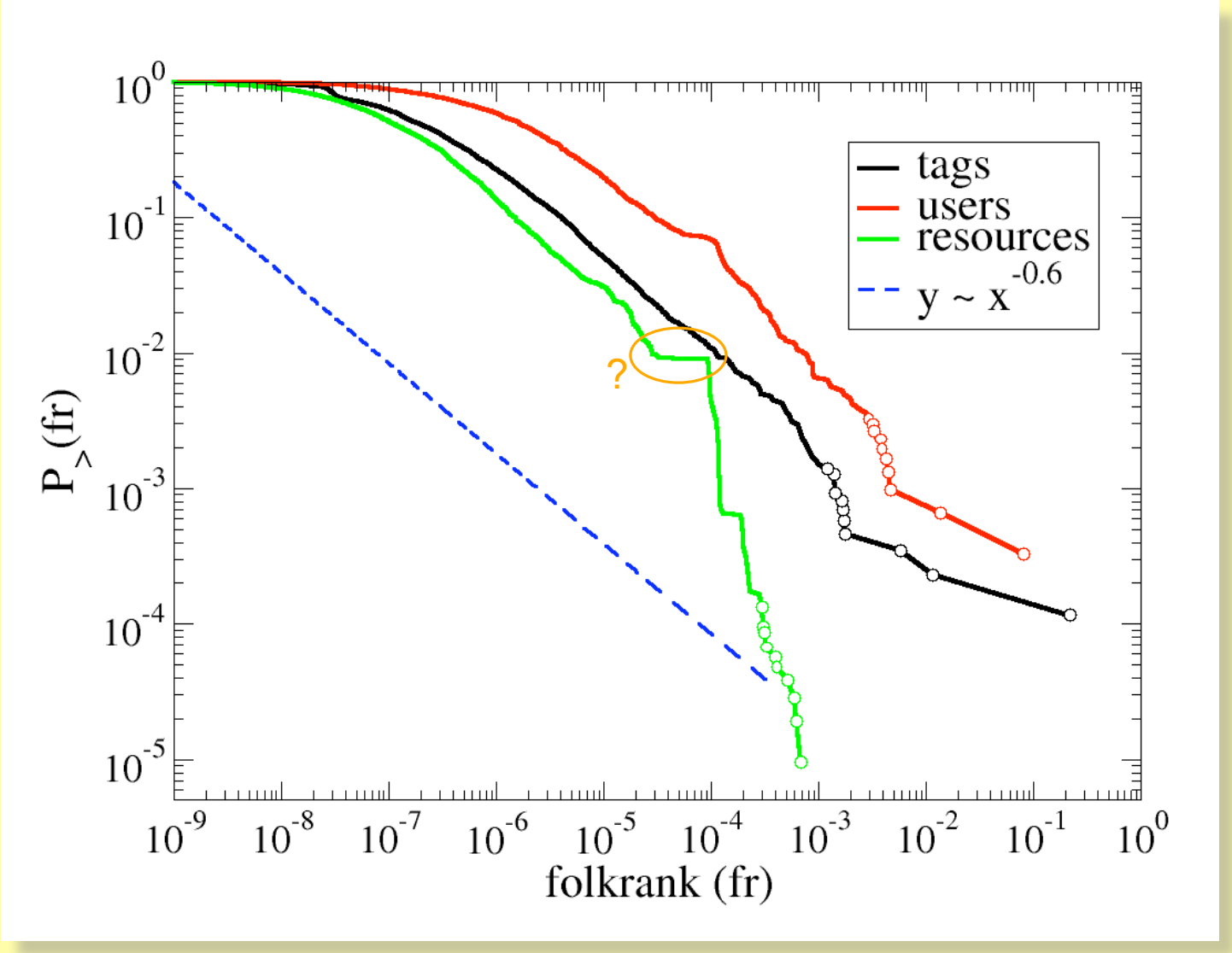
- Results for tag: **science-fiction**

BASELINE (PageRank)		PREFERENTIAL		FOLKRANK	
blog	0.011809846	sciencefiction	0.218912824	sciencefiction	0.218855247
music	0.011307161	lesezeichen_sommer_2003	0.011782188	lesezeichen_sommer_2003	0.011660334
web	0.010077613	blog	0.008920223	sf	0.005869562
design	0.009912541	music	0.008357232	diverses_und_chaotisches	0.001770573
software	0.009484339	web	0.007903793	zukunft	0.001736983
programming	0.008892541	design	0.00745063	magisterarbeit	0.001725315
linux	0.006761464	programming	0.007123381	cyberculture	0.001678076
css	0.006742816	software	0.007092474	scifi	0.001464554
news	0.006441021	sf	0.006244209	reviews	0.001433616
art	0.006345402	art	0.005410923	rezensionen	0.001431468
reference	0.006154841	reference	0.005249646	soziologie_und_psychologie	0.001418239
blogs	0.006009328	css	0.004973381	writing	0.001228814
politics	0.005759638	news	0.004934699	books	0.001028328
tools	0.005632843	linux	0.004677645	tuak	0.000942428
java	0.004929159	books	0.004607657	fiction	0.000875067
games	0.004497811	blogs	0.004602789	informatik	0.000808497
photography	0.004298107	politics	0.004599932	texte	0.000802848
mac	0.004216214	tools	0.004473632	raumfahrt	0.000771901
fun	0.004037195	java	0.003527007	weltraum	0.000752851
javascript	0.003856581	games	0.003469062	politik	0.000721332

Ranking results:

- Tagcloud [size of font prop. to log(FolkRank)]

Tags, user and resources with FolkRank > fr



MCL algorithm

- **Single iteration**:
  1. Square the normal matrix A ($M=A^2$) of the graph (two steps walk)
  2. Apply the inflation operator in order to cut away less probable paths

repeat until convergence

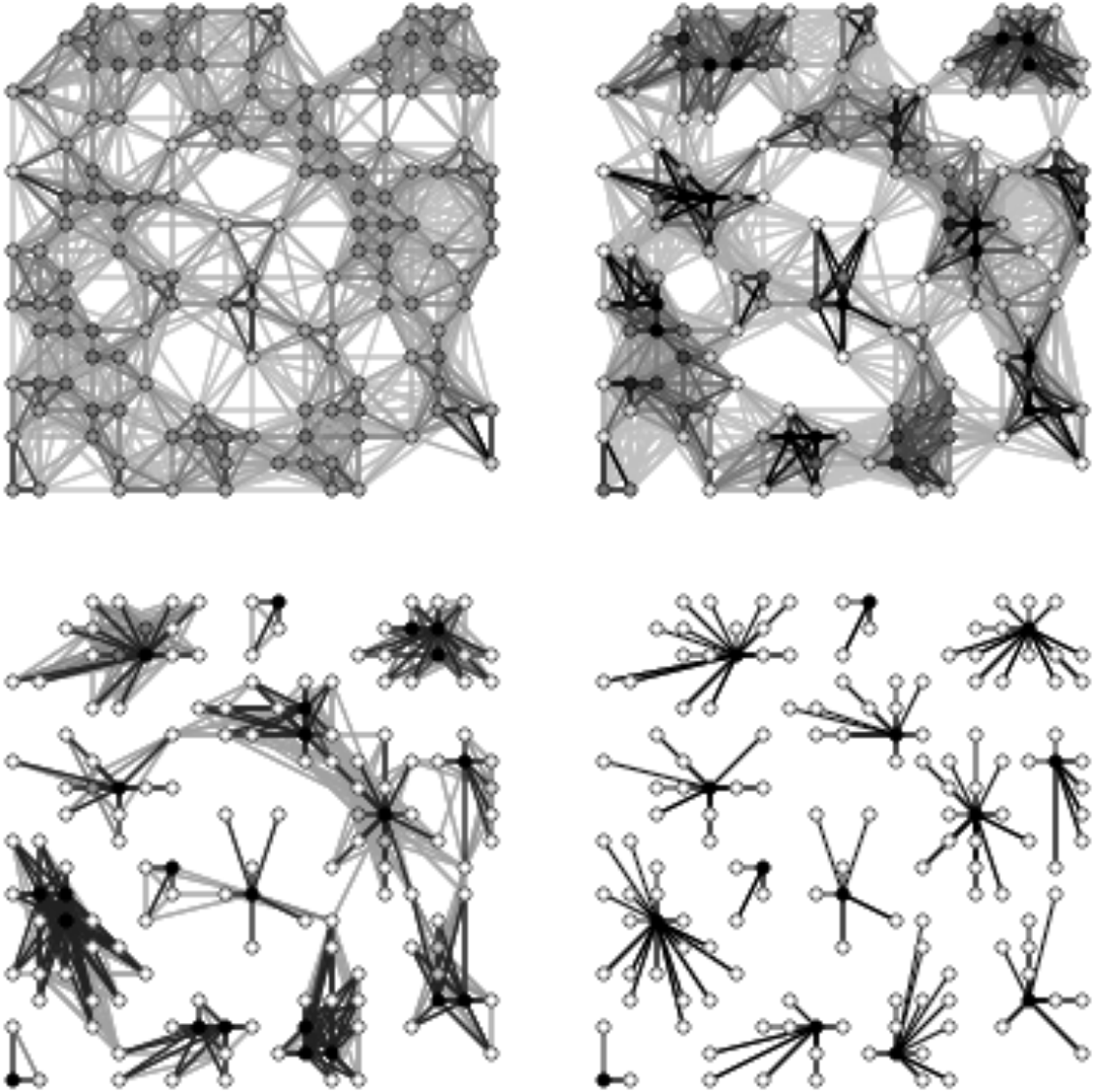
- **Inflation operator**:
  - i. Raise *each element* of the squared normal matrix to power *r>1*
  - ii. Normalize by column to recover a transition matrix

$$(\Gamma_r M)_{pq} = (M_{pq})^r / \sum_{i=1}^{k} (M_{iq})^r$$

we used r=2

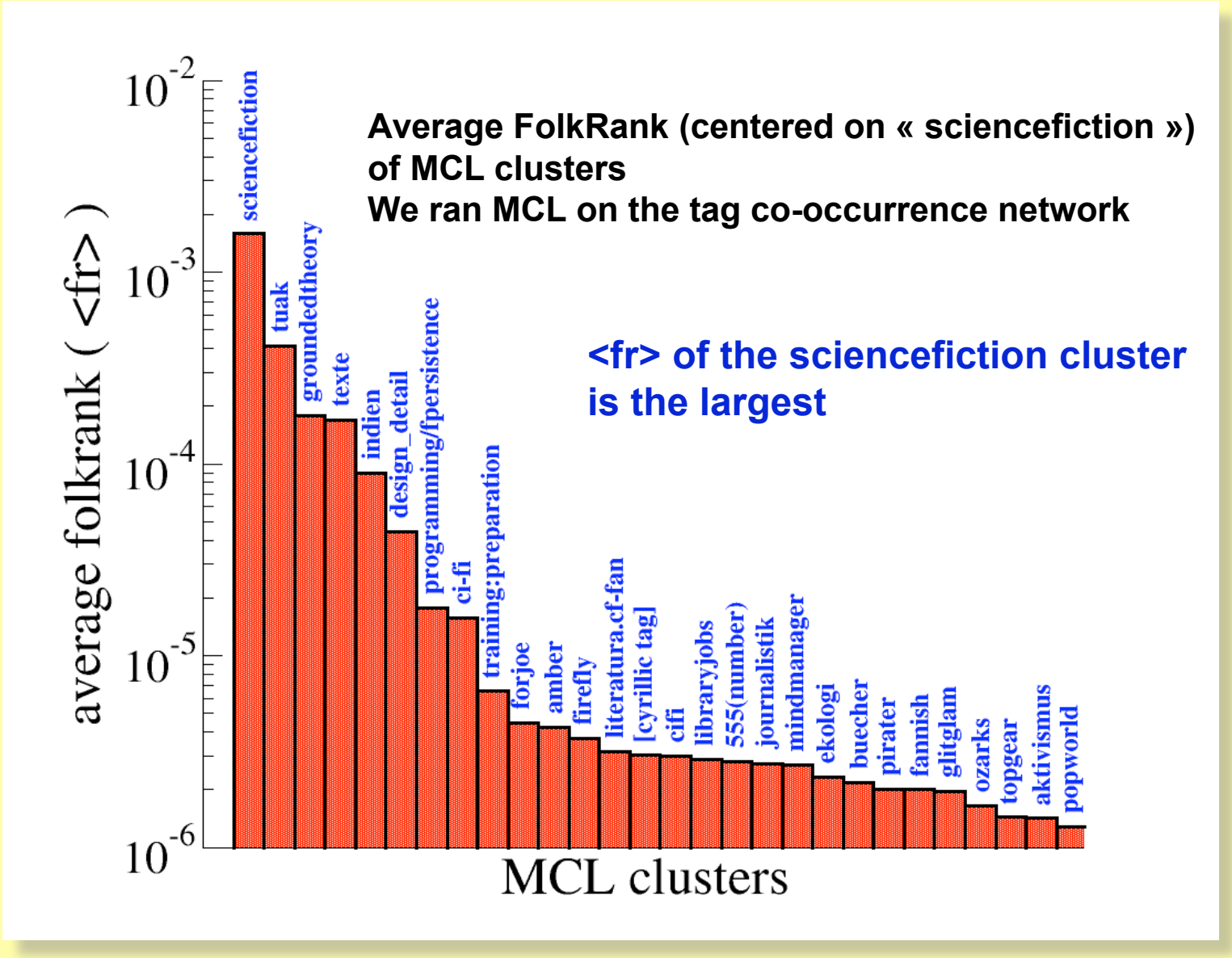
Markov Clustering Algorithm (MCL)

Stijn van Dongen,
***Graph Clustering by Flow Simulation*.**
PhD thesis,
University of Utrecht,
May 2000.

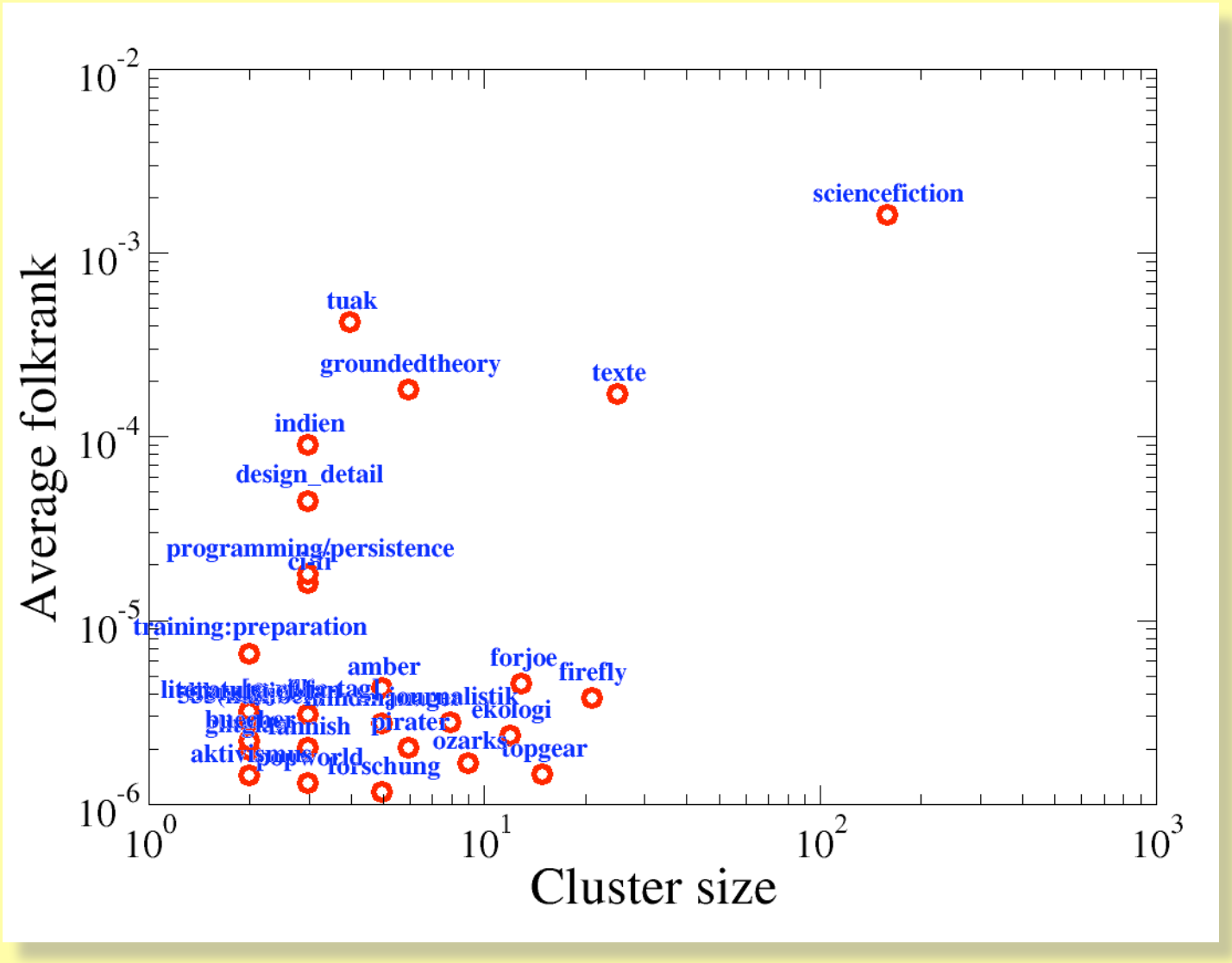


http://micans.org/mcl/

Comparison MCL vs FolkRank



Cluster size vs average FolkRank (centered on « sciencefiction ») inside the cluster



Inside the MCL cluster

sciencefiction lesezeichen_sommer_2003
diverses_und_chaotisches zukunft cyberculture rezensionen
soziologie_und_psychologie informatik raumfahrt weltraum politik
soziologie wahlen spekulatives krimskrams parteien
informatik_und_gesellschaft studienarbeit douglas_n._adams
ursula_k._leguin grüne friends_and_family bots paganism kunst fashionables
computer_und_zubehör netzwerk_neue_medien william_gibson iain_m._banks
virtuelle_nationen sozionik html_java_et_al persönliche_seiten filme ki
technologieentwicklung online-magazine zukunftsforschung cyberkultur
bruce_sterling zeitung_und_radio offiziell grüne_jugend php-resourcen jasss
niederlande online-partizipation palmpilot baden-württemberg netzpolitik junge-welt-serie
bundesverband bundestag till_westermayer avalanche iig-telematik wald
bundestagswahl_2005 interaktiv negatives viridian computergrafik arcata freund wahl
sozialesoftware weltall spd gegner zeitung/zeitschrift kifgsw retro-futurism anmeldung ökologie umfrage
news/item kritik kritiken krimis besprechungen kritiker empfehlungen krimi merkel gruene schweiz neuronal_network
wahlblog brd anarkism virtualreview neuwahlen wahlen05 cdu mypublication konst bundestagswahl anwendung kriskrug

Conclusions

- Folksonomies are interesting systems for statistical physicists
- Clustering or ranking?
  - Both approaches seem quite successful
  - MCL and FR give similar results, but:
    - MCL delivers no useful ranking
    - MCL is affected by frequency bias
    - FR differential approach seems avoid most frequency bias
  - FR deserves further theoretical study

TAGora Consortium

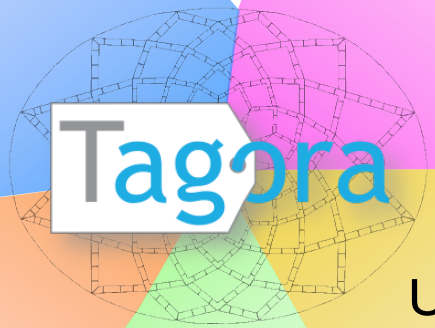
Semiotic Dynamics in Online Social Communities

University of Roma "La Sapienza"

- complex systems expertise
- modeling and simulation
- statistical tools for data analysis
- stochastic models of tagging behavior
- tag propagation via similarity metrics
- project coordination

SONY CSL

- tag-based image navigation system
- tag-based music navigation system
- feature-enhanced navigation maps
- expertise in semiotic dynamics



University of Kassel

- tagging system for bibliographic data
- data mining for folksonomies
- social network analysis

University of Koblenz-Landau

- peer-to-peer testbed for folksonomies
- representation of folksonomy data
- ontology learning

University of Southampton

- cross-folksonomy analysis
- semantic network analysis

http://www.tagora-project.eu

END

Results: adapted PageRank

http://slashdot.org/	0,0002613
http://pchere.blogspot.com/2005/02/absolutely-delicious-complete-tool.html	0,0002320
http://script.aculo.us/	0,0001770
http://www.adaptivepath.com/publications/essays/archives/000385.php	0,0001654
http://johnvey.com/features/deliciousdirector/	0,0001593
http://en.wikipedia.org/wiki/Main_Page	0,0001407
http://www.flickr.com/	0,0001376
http://www.goodfonts.org/	0,0001349
http://www.43folders.com/	0,0001160
http://www.csszengarden.com/	0,0001149
http://wellstyled.com/tools/colorscheme2/index-en.html	0,0001108
http://pro.html.it/esempio/nifty/	0,0001070
http://www.alistapart.com/	0,0001059
http://postsecret.blogspot.com/	0,0001058
http://www.beelerspace.com/index.php?p=890	0,0001035
http://www.techsupportalert.com/best_46_free_utilities.htm	0,0001034
http://www.alvit.de/web-dev/	0,0001020
http://www.technorati.com/	0,0001015
http://www.lifehacker.com/	0,0001009
http://www.lucazappa.com/brilliantMaker/buttonImage.php	0,0000992
http://www.engadget.com/	0,0000984

Results: adapted PageRank

Tag	ad. PageRank
system:unfiled	0,0078404
web	0,0044031
blog	0,0042003
design	0,0041828
software	0,0038904
music	0,0037273
programming	0,0037100
css	0,0030766
reference	0,0026019
linux	0,0024779
tools	0,0024147
news	0,0023611
art	0,0023358
blogs	0,0021035
politics	0,0019371
java	0,0018757
javascript	0,0017610
mac	0,0017252
games	0,0015801
photography	0,0015469
fun	0,0015296

User	ad. PageRank
shankar	0,0007389
notmuch	0,0007379
fritz	0,0006796
ubi.quito.us	0,0006171
weev	0,0005044
kof2002	0,0004885
ukquake	0,0004844
gearhead	0,0004820
angusf	0,0004797
johncollins	0,0004668
mshook	0,0004556
frizzlebiscuit	0,0004543
rafaspol	0,0004535
xiombarg	0,0004520
tidesonar02	0,0004355
cyrusnews	0,0003829
bldurling	0,0003727
onpause_tv_anytime	0,0003600
cataracte	0,0003462
triple_entendre	0,0003419
kayodeok	0,0003407

Results: boomerang

Preference for tag: boomerang

PageRank without preference

Tag	ad. PageRank
system:unfiled	0,0078404
web	0,0044031
blog	0,0042003
design	0,0041828
software	0,0038904
music	0,0037273
programming	0,0037100
css	0,0030766
reference	0,0026019
linux	0,0024779
tools	0,0024147
news	0,0023611
art	0,0023358
blogs	0,0021035
politics	0,0019371
java	0,0018757
javascript	0,0017610
mac	0,0017252
games	0,0015801
photography	0,0015469
fun	0,0015296

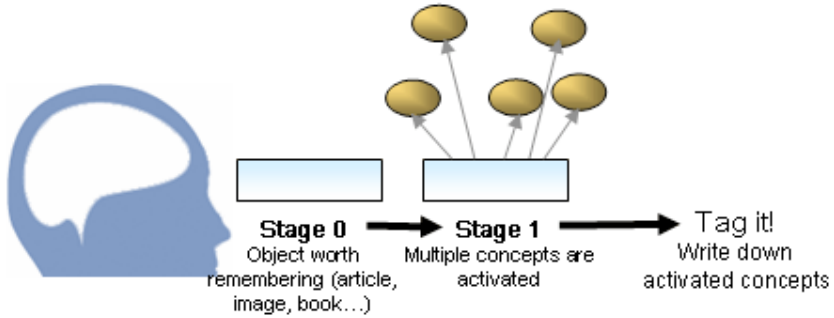
PageRank with preference

Tag	ad. PRank
boomerang	0,4036883
shop	0,0069058
lang:de	0,0050943
software	0,0016797
java	0,0016389
programming	0,0016296
web	0,0016043
reference	0,0014713
system:unfiled	0,0014199
wood	0,0012378
kassel	0,0011969
linux	0,0011442
construction	0,0011023
plans	0,0010226
network	0,0009460
rdf	0,0008506
css	0,0008266
design	0,0008248
delicious	0,0008097
injuries	0,0008087
pitching	0,0007999

FolkRank with preference

Tag	FolkRank
boomerang	0,4036867
shop	0,0066477
lang:de	0,0050860
wood	0,0012236
kassel	0,0011964
construction	0,0010828
plans	0,0010085
injuries	0,0008078
pitching	0,0007982
rdf	0,0006619
semantic	0,0006533
material	0,0006279
trifly	0,0005691
network	0,0005568
webring	0,0005552
sna	0,0005073
socialnetworkanalysis	0,0004822
cinema	0,0004726
erie	0,0004525
riparian	0,0004467
erosion	0,0004425

“Social Computing”

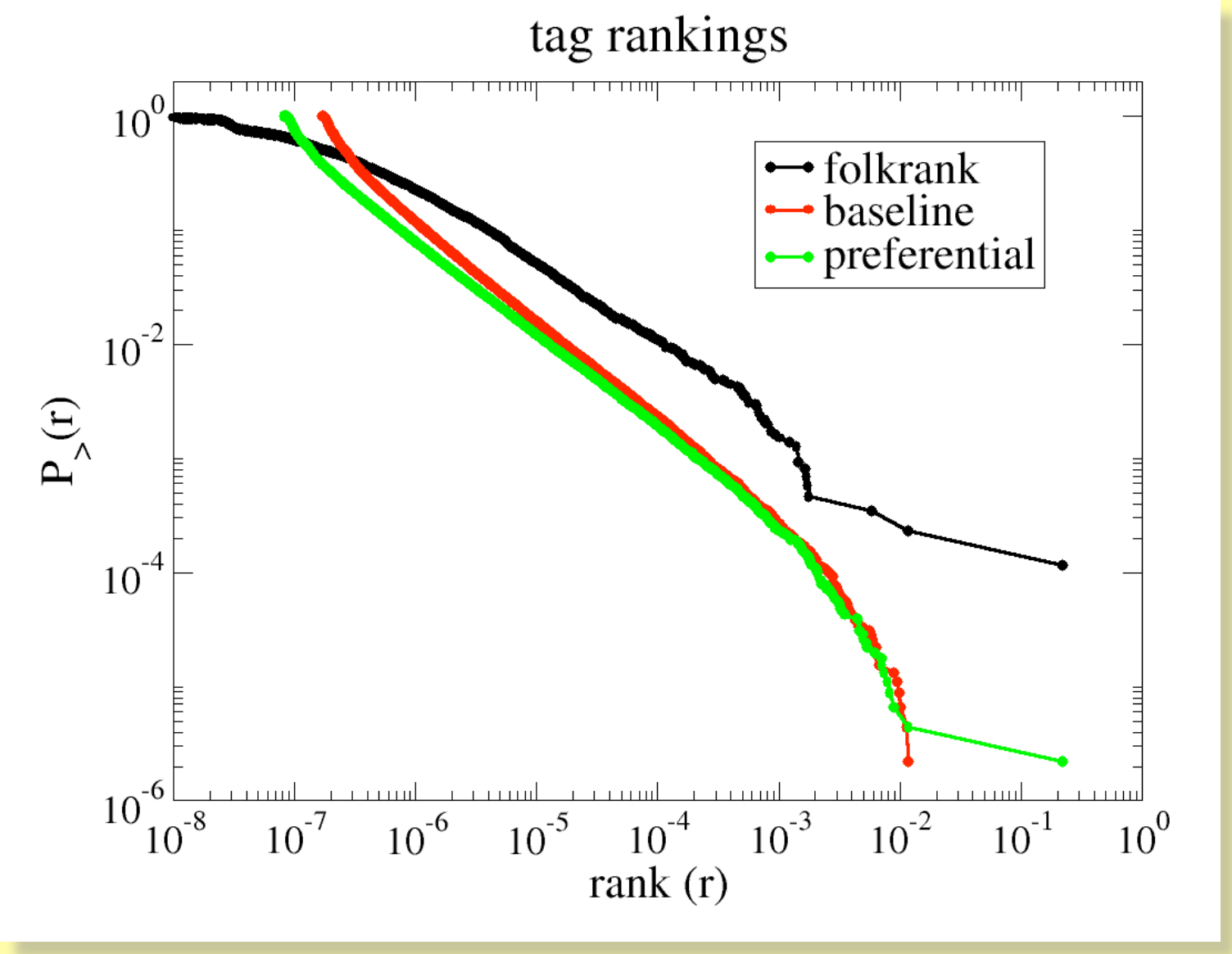


- online communities
- media-centric
- web-based interaction
- small cognitive overhead

- collaborative tagging and folksonomies
- online collaborative gaming
- collaborative filtering
- recommendation/trust networks

http://www.peekaboom.o

(normalized) nr of tags with Folk-, Page-, Preferential-Rank greater than *r*



Best ten tags from FolkRank centered on « Science Fiction »