

Unless otherwise noted, the content of this course material is licensed under a Creative Commons Attribution 3.0 License.

<http://creativecommons.org/licenses/by/3.0/>

Copyright 2008, Lada Adamic

You assume all responsibility for use and potential liability associated with any use of the material. Material contains copyrighted content, used in accordance with U.S. law. Copyright holders of content included in this material should contact [open.michigan@umich.edu](mailto:open.michigan@umich.edu) with any questions, corrections, or clarifications regarding the use of content. The Regents of the University of Michigan do not license the use of third party content posted to this site unless such a license is specifically granted in connection with particular content objects. Users of content are responsible for their compliance with applicable law. Mention of specific products in this recording solely represents the opinion of the speaker and does not represent an endorsement by the University of Michigan. For more information about how to cite these materials visit <http://michigan.educommons.net/about/terms-of-use>.



School of Information  
University of Michigan

# Collaborative filtering & tagging networks

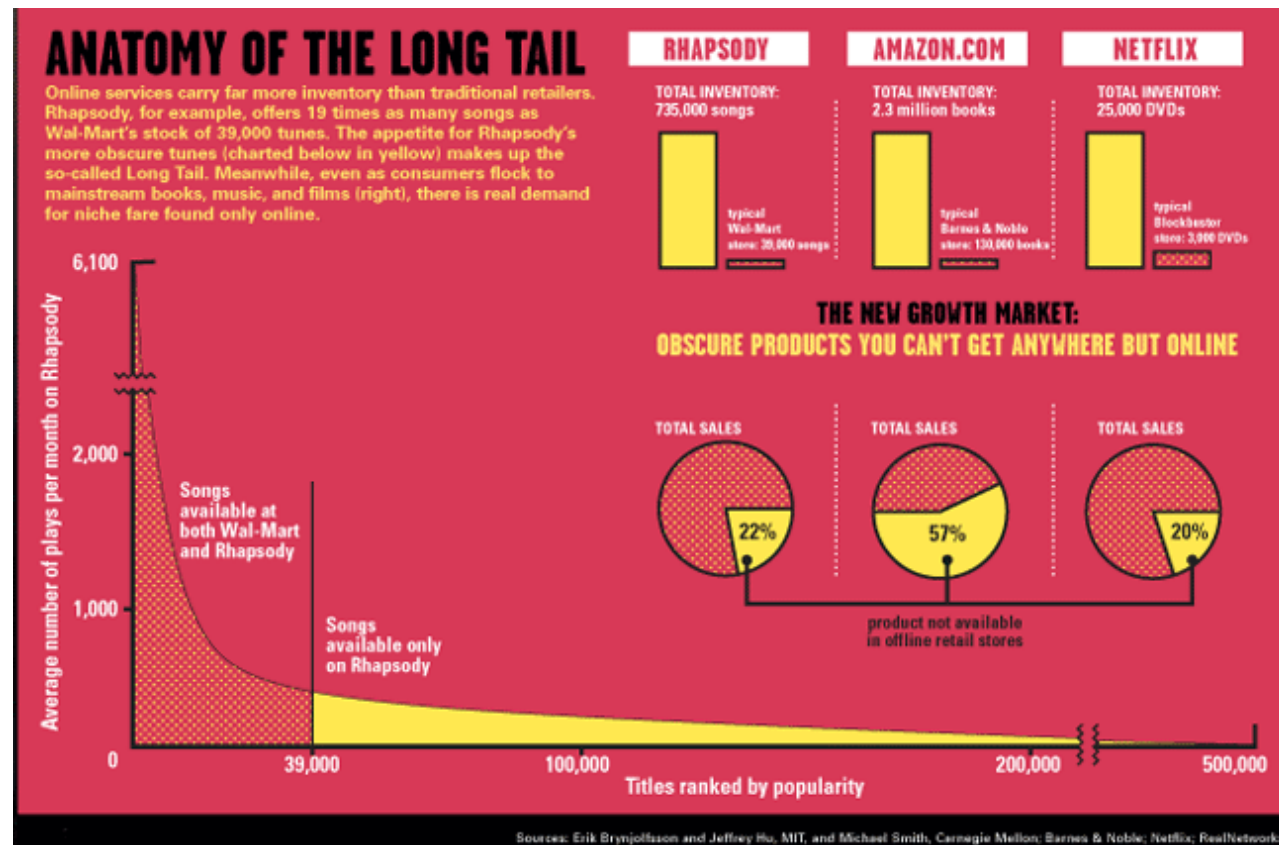


## outline

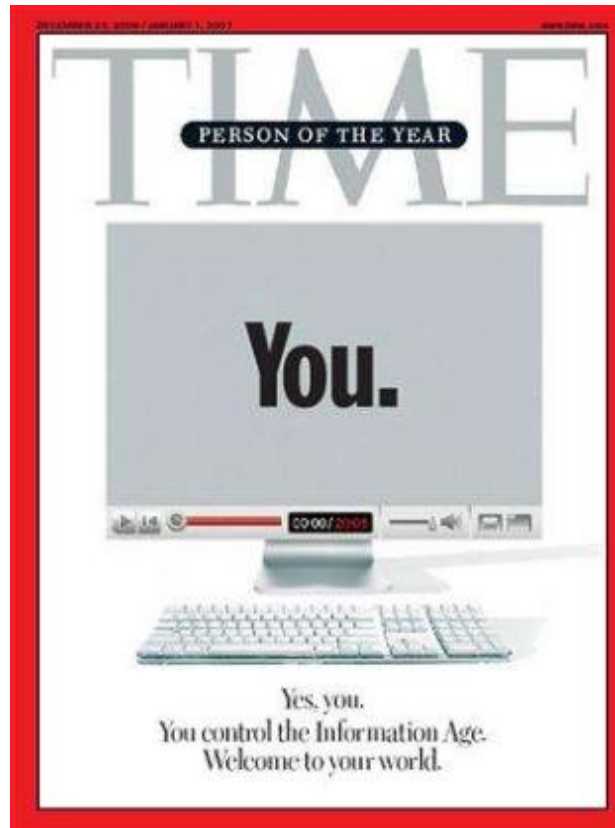
- motivation for collaborative filtering
  - the Long Tail of content popularity
  - unprecedented amount of user-generated content
- tagging as a tripartite network/hypergraph
  - evolution of the tagging network
- pitfalls of collaborative tagging

# The Long Tail

- The internet enables the distribution of niche items
- Need a way to discover items that match our interests & tastes among tens or hundreds of thousands



Chris Anderson, 'The Long Tail', Wired, [Issue 12.10](#) - October 2004



- That is you (plural) not you (singular)!
- Collaborative content tagging, and filtering is allowing the little guys (like you and me) to find audience for and discover new content

Source: <http://www.time.com/time/covers/0,16641,20061225,00.html>

## when people search alone...

query	count	
how to tie a tie	92	
how to		58
how to write a resume	47	
how to have sex	25	
how to lose weight		23
how to build a deck	23	
how to get pregnant	21	
how to write a bibliography	20	
how to gain weight	19	
how to kiss	18	
how to get a Passport	17	
how to write a cover letter	17	
How to lose a guy in 10 days		17
how to draw	14	
how to pass a drug test		14
how to knit		13
how to write a book		13
how to ask for a raise		13
how to play guitar	13	
how to save money	13	
how to play poker	12	
how to get rid of ants	12	
how to start a business		11
how to make money	11	
how to draw anime	11	
how to draw manga	11	
how to pray the rosary	10	



# You're Invited, San Francisco Bay Area

What is TiVo?

Buy TiVo

Setup & Support

I Have TiVo!

TiVo Central Online

Rewards

TiVo Fan!

Manage my account

Exclusive TiVo service features

TiVo community

Developers

Tips & tricks

Showcases

About TiVo Inc.

Online Scheduling

TiVo Rewards

Manage My Account

Careers

Contact Us

Customer Service

Activate or Upgrade TiVo Service.

Developers

## TiVo is throwing a singles' mixer!



Ever wish your TiVo® WishList® or TiVo Suggestions could score YOU the perfect match? Come flirt with the possibility of finding your own special someone, "TiVo-style." PLUS get 2 free drinks AND be automatically entered in a raffle for one of 14 brand-new TiVo boxes with product lifetime subscription! Must be present to win.

Already found the love of your life? Bring a single friend — or just forward this email — and simply share your love for the amazing TiVo service. We've got lots to talk about!

**VERY LIMITED GUEST LIST! RSVP before it's too late!**

**Where:** A bar in the hippest downtown San Francisco hotel (We'll tell you later!)

**When:** Monday, February 13, 2006

**Time:** Registration begins at 6:30 (come right after work!); Party begins at 7:00 pm

**Why:** You can tell a lot about a person from the TV shows they watch! Let your Now Playing list be your guide.

**How:** RSVP by taking our [TiVo MatchMaking Quiz](#).

**Cost:** FREE! PLUS 2 FREE drinks on TiVo. You'll be entered in a raffle for one of 14 TiVo boxes with product lifetime subscription, so you can watch your favorite shows with the ones YOU love!

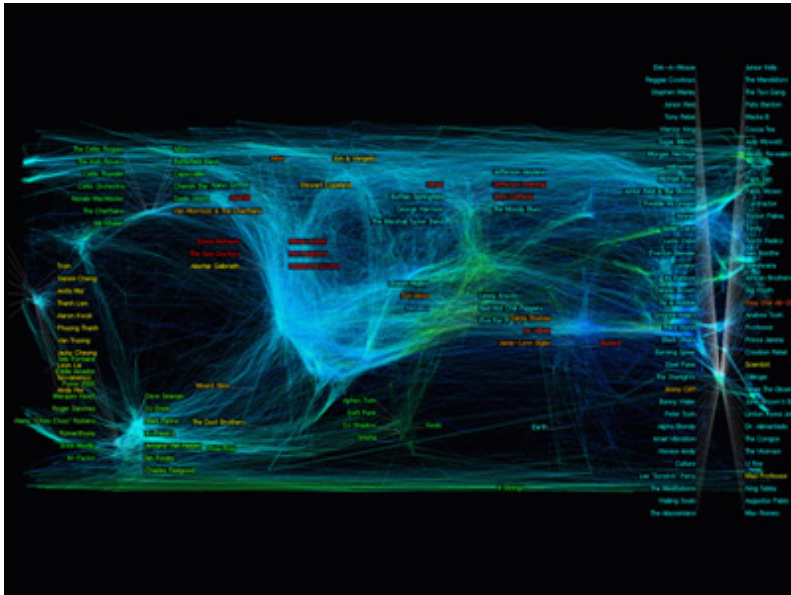
**Sorry. Due to the overwhelming Love for TiVo in the San Francisco Bay Area, our guest list for the TiVo Singles Mixer is now closed.**

From all of us at TiVo, here's to finding true love on Valentine's Day!

For questions about this event, please see our [rules and regulations](#).

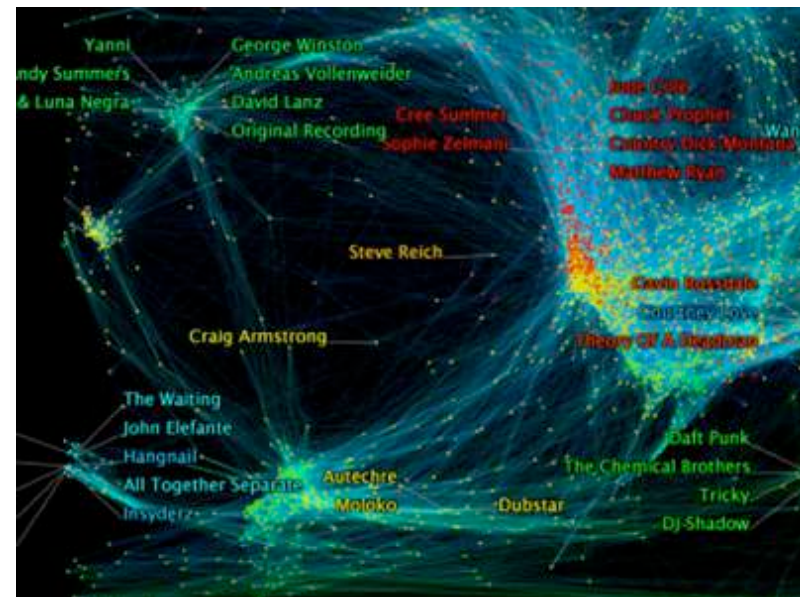
Source: Tivo, <http://www.tivo.com/>

## Example: Yahoo music recommends similar songs/artists



- By rating and listening to music you let Y! Music know your tastes
- Y! Music customizes suggestions/ radio station to match your taste
- Demo this interactive graph at:  
<http://www.stanford.edu/~dgleich/demos/worldofmusic/interact.html>

- Instant message what you're playing to friends
- service suggests 'influencers' who match your taste
- you can choose your own influencers...

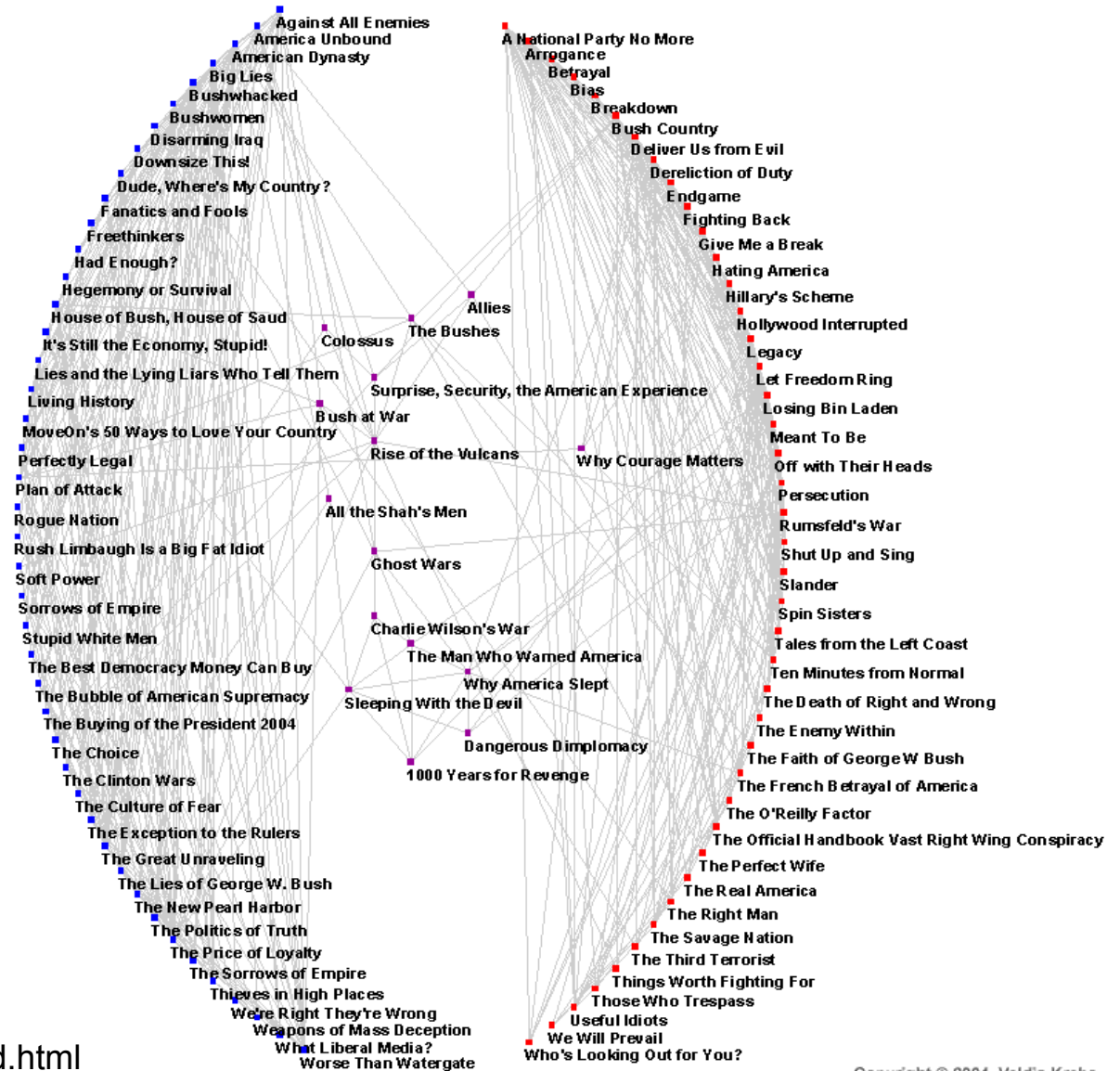


Source: M. R. David Gleich, Matt Rasmussen, Leonid Zhukov and K. Lang. The World of Music: SDP layout of high dimensional data. In Info Vis, 2005.



# Collaborative filtering and polarized topics

- It seems to work for book topics
- Valdis Krebs, “political books”



■ <http://www.orgnet.com/divided.html>

## Recommendations: document centric view

- Knowledge brief viewer at HP
  - find documents, experts, and other readers related to the document




- Not necessarily something you want to do on Amazon, but within a well defined and technical space

Source: intranet screenshot, HP

## Recommendations: user centric view

- find others like you based on your writing/download history




**Paul**



Tech Consulting   Systems Integration

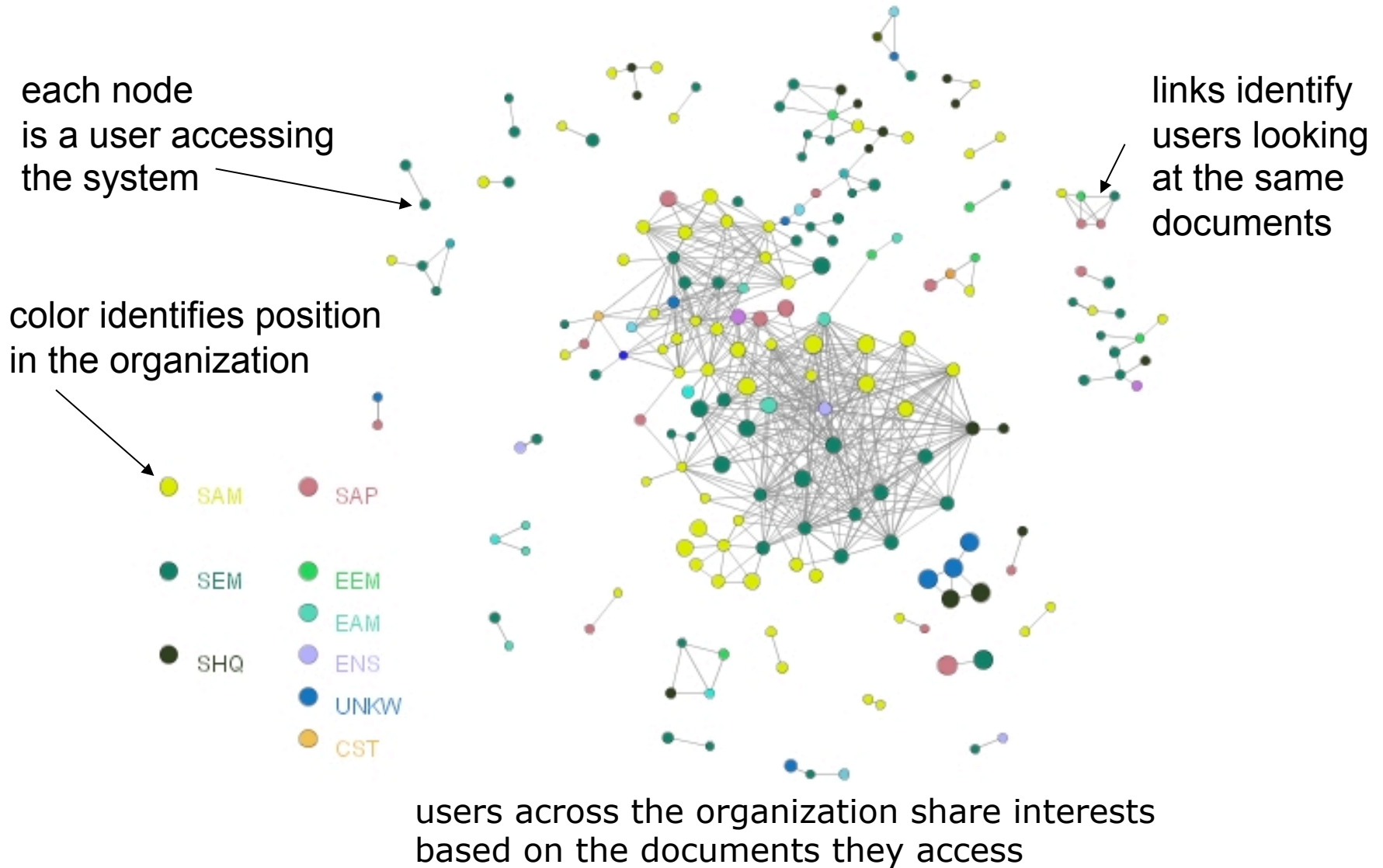
32 docs viewed

Paul is a consultant with the .NET Solutions group within the Practice in Minneapolis, Minnesota. Paul specializes in e-commerce UI and middle tier development and their related Microsoft technologies. In his spare time he enjoys the freezing Minnesota weather, cheering for the Vikings, Twins, Wolves and Wild and traveling the world.

users similar to Paul Johansen						
sim	name	unit	group	function	family	#docs
0.35	 <a href="#">John</a>			Solution Architech	Systems Integration	30
		<p>John is a member of the .NET Results North American Team. He has extensive experience developing customized solutions in Domino, Microsoft, and WebSphere. He is certified MCSD for .NET, MCAD for .NET, MCSD for Visual Studio 6.0, MCSE for Windows 2000, and MCDBA for MSSQL 2000.</p>				
0.29	 <a href="#">Tom</a>			Tech Consulting	Systems Integration	236
		<p>Tom is a consultant for the Enterprise Microsoft Services .Net Solutions practice. Tom has worked on a variety of custom software projects based on Microsoft technologies.</p>				
0.26	 <a href="#">Martyn</a>	SEM	EMCI	Tech Consulting	Systems Integration	46
		<p>Martyn is a member of EMEA C&amp;I currently working with Microsoft .NET. He has been designing, developing, and testing various kinds of software since 1979 and has experienced many examples of "how not to do things". He has worked on many projects and is experienced in the full project lifecycle. His current interests are round all things .Net.</p>				

Source: Lada Adamic

# Mapping knowledge communities from download patterns



another example of expertise search,  
this time using occurrence of names in publicly available documents

## PeopleFinder<sup>2</sup>



Search by: [Person](#) [Department](#) [Topic](#)

[Advanced Search](#)

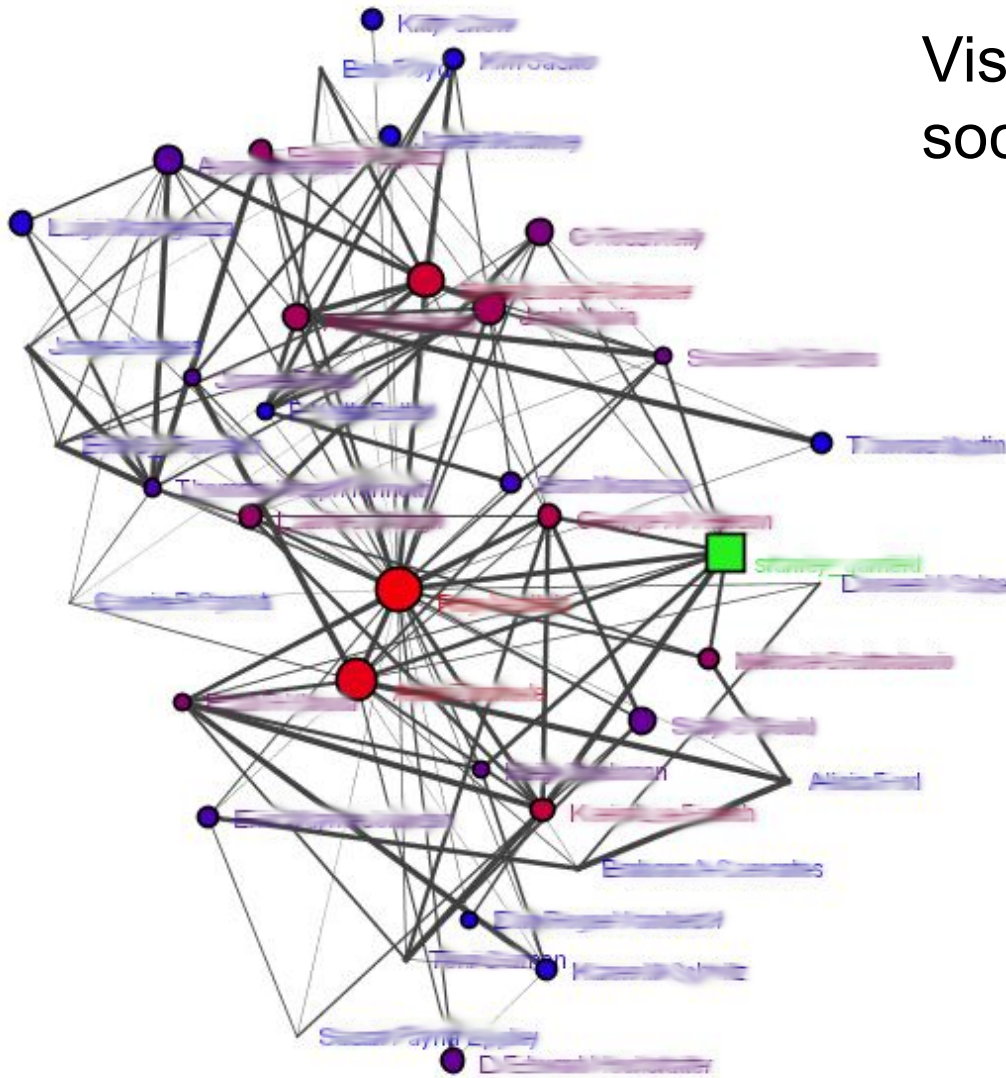
- [Help?](#)
- [What's new?](#)

A short technical description is now available [here](#)

## People associated with *"knowledge management"*

enter your SEA (e.g. "joe.schmoe@hp.com") to see how you can connect to these people

Score	Name
<a href="#">100.00</a>	<a href="#">Bob</a> (dept1) <a href="#">See matches...</a>
<a href="#">57.14</a>	<a href="#">Sam</a> (dept2) <a href="#">See matches...</a>
<a href="#">42.85</a>	<a href="#">Kate</a> (dept3) <a href="#">See matches...</a>
<a href="#">42.85</a>	<a href="#">Harold</a> (dept4) <a href="#">See matches...</a>
<a href="#">28.57</a>	<a href="#">Richard</a> (dept2) <a href="#">See matches...</a>
<a href="#">28.57</a>	<a href="#">Uwe</a> (dept5) <a href="#">See matches...</a>



Visualize a person's social network

## Using social network information to suggest how you may be connected to experts

using 'lada.adamic@hp.com' as the starting user (it's true, I've collaborated with "Victor")

Score	Name
<a href="#">100.00</a>	<a href="#">Bob (dept1)</a>
	■ <a href="#">See matches...</a>
	72 Lada Adamic > Victor > Bob
<a href="#">57.14</a>	<a href="#">Sam (dept2)</a>
	■ <a href="#">See matches...</a>
	71 Lada Adamic > Victor > Sam
<a href="#">42.85</a>	<a href="#">Kate (dept3)</a>
	■ <a href="#">See matches...</a>
	59 Lada Adamic > Victor > Bob > Kate
	56 Lada Adamic > Victor > Sam > Kate
<a href="#">42.85</a>	<a href="#">Harold (dept4)</a>
	■ <a href="#">See matches...</a>
	57 Lada Adamic > Victor > Bob > Harold



**You get copied on an email that has a number of people you have and have not worked with. Who are they? Where are they in the organization? How do you connect to them? How do they connect to each other? Do they work with other people that you have heard of?**

### Hierarchy structure

Keith > M3 > M1 > Carly  
Joe > M7 > M6 > Keith > M3 > M1 > Carly  
Sam > M8 > Richard > Tim > M2 > Carly  
Luke > M4 > M2 > Carly  
Marion > Toby > M1 > Carly

### Interconnections

Lada and Sam

->

**Victor**

Luke and Sam

->

Richard, Simon, Chris, Philip, Sophie, Allen, Susan, **Craig**, Tim, Bob, Mark, Kate and **Toby**

## **Social tagging** **a method of explicit social search**

- More than just like or dislike, download or not
  - categorize & comment
- folksonomies: users collectively label items which can then be retrieved by others
  
- Blogs, del.icio.us (and other social bookmarking systems e.g. CiteULike), Flickr
- digg (alternative to slashdot for techie news)
  - “With digg, users submit stories for review, but rather than allow an editor to decide which stories go on the homepage, the users do.”

## latest front page stories

200

diggs

[Finally: Official DS vs DS Lite comparison pictures! \(including brightness\)](#) submitted by [DevilsRejection](#) 14 hours 9 minutes ago (via <http://www.genmay.com/showthre...>)

No more concepts, no more CGI, someone who went to the DS Conference in Japan yesterday (Feb, 15, 2006)

[digg it](#)[24 comments](#) | [blog this](#) | [email this](#) | category: [gaming](#)

425

diggs

[Students at MIT give flying car a shot](#) submitted by [fuzzyjit](#) 15 hours 52 minutes ago (via <http://news.com.com/StudentsCa...>)

An SUV with retractable wings could make 100- to 500-mile jumps and carry two people and luggage on a single tank of gas.

[digg it](#)[35 comments](#) | [blog this](#) | [email this](#) | category: [science](#)

244

diggs


[iLife '06 Updates Available](#) submitted by [snipehack](#) 15 hours 28 minutes ago (via <http://www.macrumors.com/pages...>)

Every aspect of iLife has been updated. iDVD - This update to iDVD 6 resolves issues with integration with the other iLife applications, importing of legacy projects and some theme related issues. It also addresses a number of other minor issues.

[digg it](#)[16 comments](#) | [blog this](#) | [email this](#) | category: [apple](#)

269

diggs

[Babies have an innate ability to do simple maths](#) submitted by [Bungledust](#) 1 day 1 hour ago (via [http://www.abc.net.au/science/...](http://www.abc.net.au/science/))

Even before babies learn to talk they have a bit of a grasp of maths, according to new research concluding that infants may have an abstract sense of numerical concepts.

[digg it](#)[22 comments](#) | [blog this](#) | [email this](#) | category: [science](#)

## Social tagging - Flickr

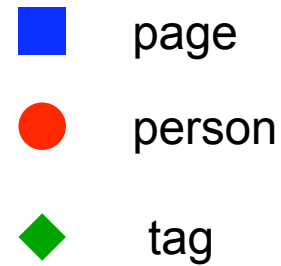
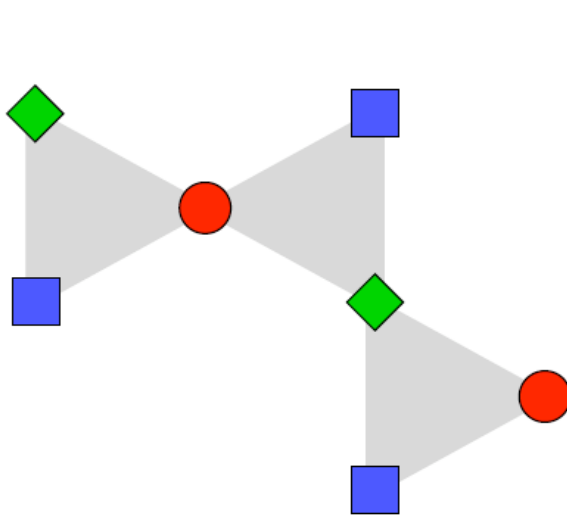


- Image search much more difficult than textual search
- solution: tagging
- One person's *nose* is another person's *cat* or *Katze*



Source: polandeze, flickr; <http://creativecommons.org/licenses/by/2.0/deed.en>

# tripartite/hypergraph tagging graphs



- Can project onto bipartite graphs
  - person – tag
  - tag – page
  - person – page
- Can project onto one-mode graphs
  - person – person
  - tag – tag
  - person - page

## Modeling the growth of tagging networks

- users become aware of popular items and tag them
- users copy others' tags
- users tend to use their own tags...

# All the little side effects of living digitally

- Find out the coolest/newest things from what people are
  - blogging, tagging, emailing, searching

## The New York Times

### Most E-Mailed

1. [A Conversation With Deborah Tannen: Author Applies Tools of Linguistics to Mend Mother-Daughter Divide](#)
2. [More and More, Favored Psychotherapy Lets Bygones Be Bygones](#)
3. [A Cancer Drug Shows Promise, at a Price That Many Can't Pay](#)
4. [Big Study Finds No Clear Benefit of Calcium Pills](#)
5. [Op-Ed Contributor: The Kiss of Life](#)

what's this?

Technorati™ Search Tags Blog Finder Explore

Search 28 million blogs for the latest on:  SEARCH

Options

Top Searches

- Brrreeport
- "Abu Ghraib"
- Chenev
- "Du Bist Deutsch..."
- "Office Live"

More in Search >>

Hot Tags

- Cheney
- Dick Cheney
- brrreeport
- Islam
- Cartoons

More in Tags >>

Featured Blog **Add Yours!**

Roachs Online Review

Tagged Attitude, buckingham, fizzy, gerald, Gig, Live music, mick...

More in Blog Finder >>

what is going on in the German blogosphere?

Source: Most E-Mailed – The New York Times, <http://www.nytimes.com>

Source: Technorati, <http://www.technorati.com>

# Brrreeeport: how long does it take for news to get around?

Blog

## brrreeeport

*Blog* : [Blog](#), posted 15-FEB-2006 10:34, by *M Freitas*

---

The object of this post is to see how long before search engines and trackers (such as [Technorati](#) and [Google Blog Search](#)) pick up new tags on the blogosphere...

So I am just adding Geekzone to the whole [brrreeeport started by Scoble](#).

The whole experiment involves creating a new word, not present in any search engine, and try to replicate it as far as possible... I know news.google.com gets Geekzone stories up to 10 minutes after publication - but what about the other tools?

*Update*: Nice, Google Blog Search found this post in less than 10 minutes, while Technorati found it in 15 minutes!

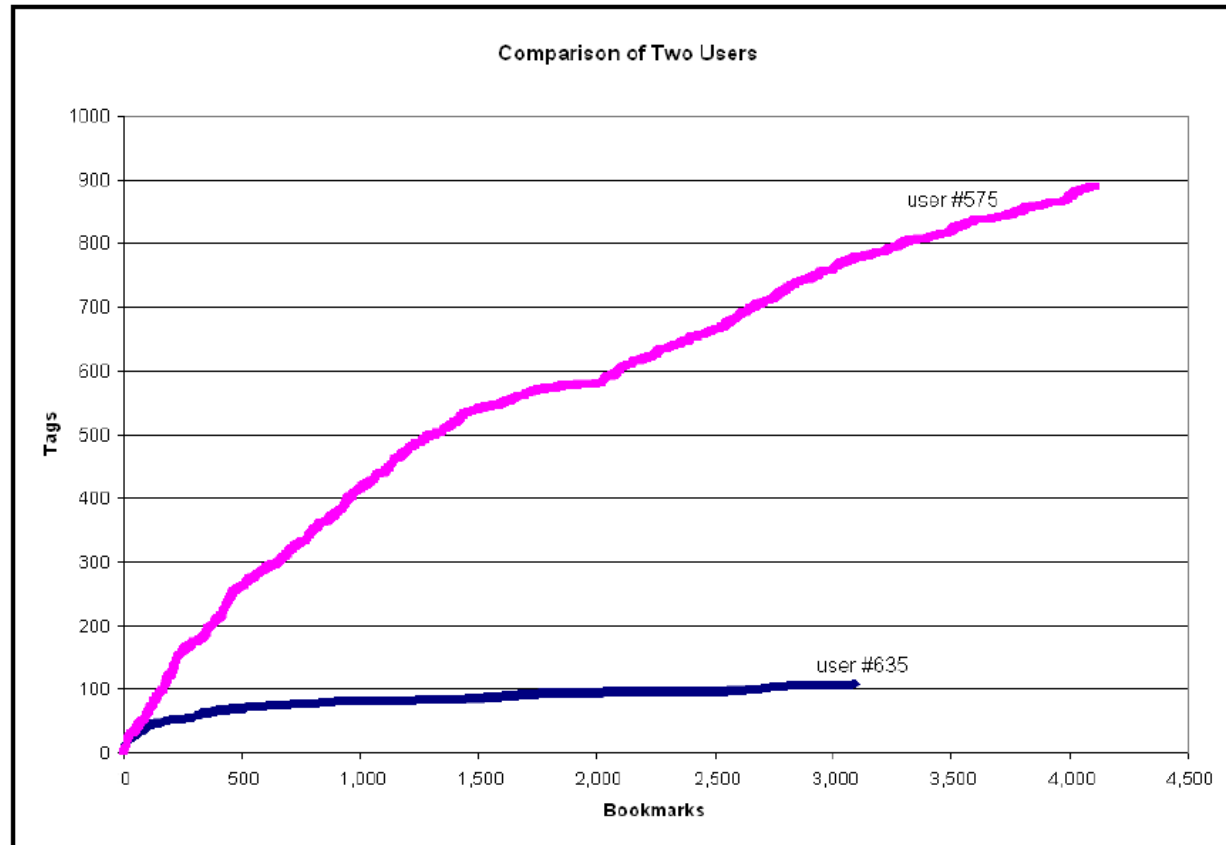
Source: M Freitas



## tag purpose – which ones are useful for social search?

- **1. Identifying What (or Who) it is About.**
  - identify topics. include common nouns, proper nouns (people or organizations).
- **2. Identifying What it Is.**
  - e.g. *article, blog and book.*
- **3. Identifying Who Owns It.**
  - e.g. a blogger
- **4. Refining Categories.**
  - e.g. numbers, especially round numbers (e.g. 25, 100)
- **5. Identifying Qualities or Characteristics.**
  - Adjectives expressing opinion such as *scary, funny, stupid...*
- **6. Self Reference.**
  - Tags beginning with “my,” like *mystuff* and *mycomments*
- **7. Task Organizing.**
  - grouping information together by task. Examples include *toread, jobsearch.*

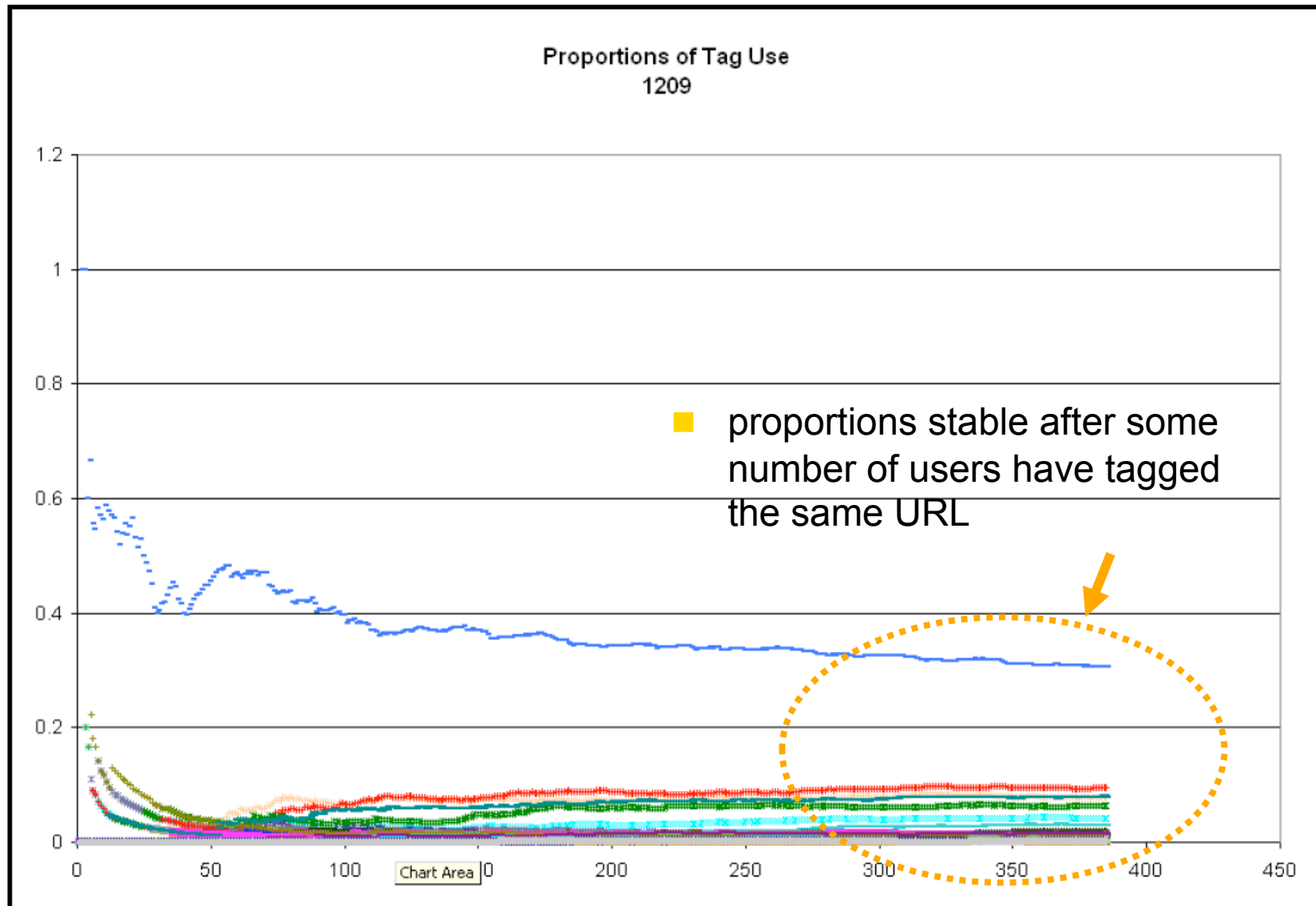
## del.icio.us (study by Golder and Huberman)



- some users use mostly same-old tags for everything, others create new ones at a fast rate

Source: Golder, S. and Huberman, B. A. (2006) Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198--208.

## tag proportions – different tags for different people?

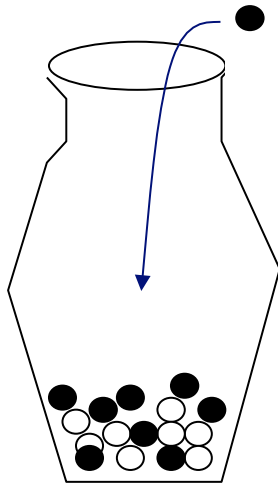


Source: Golder, S. and Huberman, B. A. (2006) Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198--208.

# simple model of user behavior

## ■ Polya's urn (contagion model)

draw a ball, note it's color replace the ball, and place another ball of the same color in the urn



del.icio.us suggests tags used by others in order of popularity

learn more about our tagometer badge on our [blog!](#) [hide this](#)

---

**del.icio.us / ladamic /**

[your bookmarks](#) | [your network](#) | [subscriptions](#) | [links for you](#) | [post](#)

[popular](#) | [recent](#)

logged in as **ladamic** | [settings](#) | [logout](#) | [help](#)

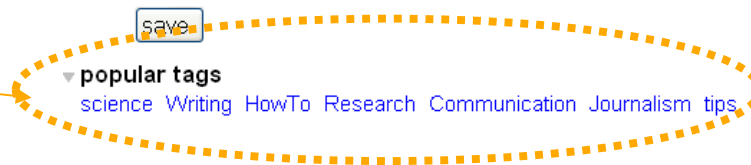
---

url

description

notes

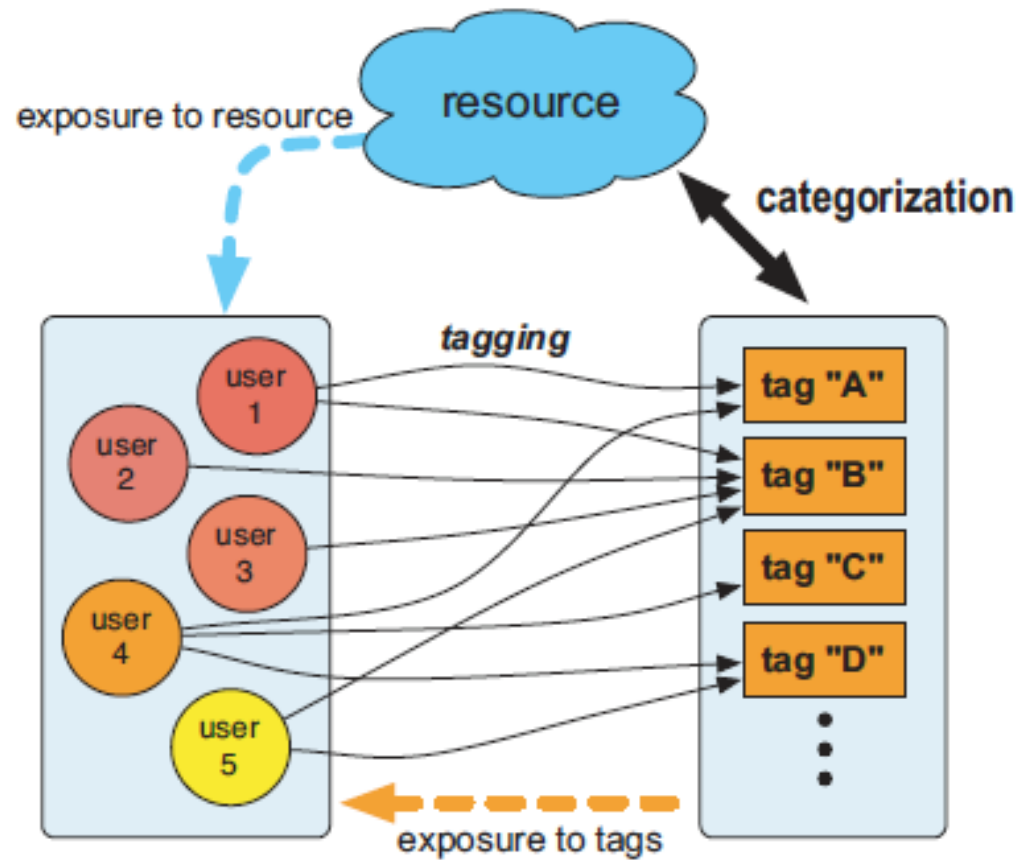
tags  space separated



Source: del.icio.us, <http://del.icio.us>

# tagging activity

- Catutto et al. PNAS 2006

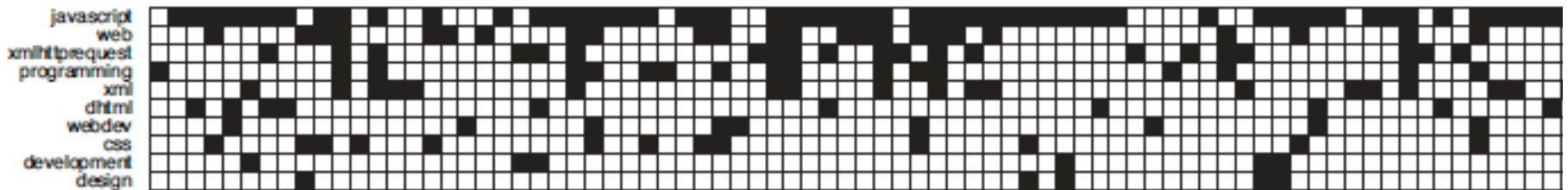
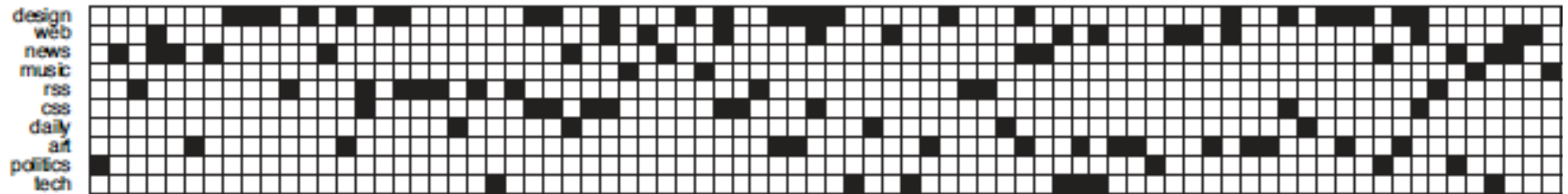


Source: Semiotic dynamics and collaborative tagging;

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1785269>

# time evolution

■ blog



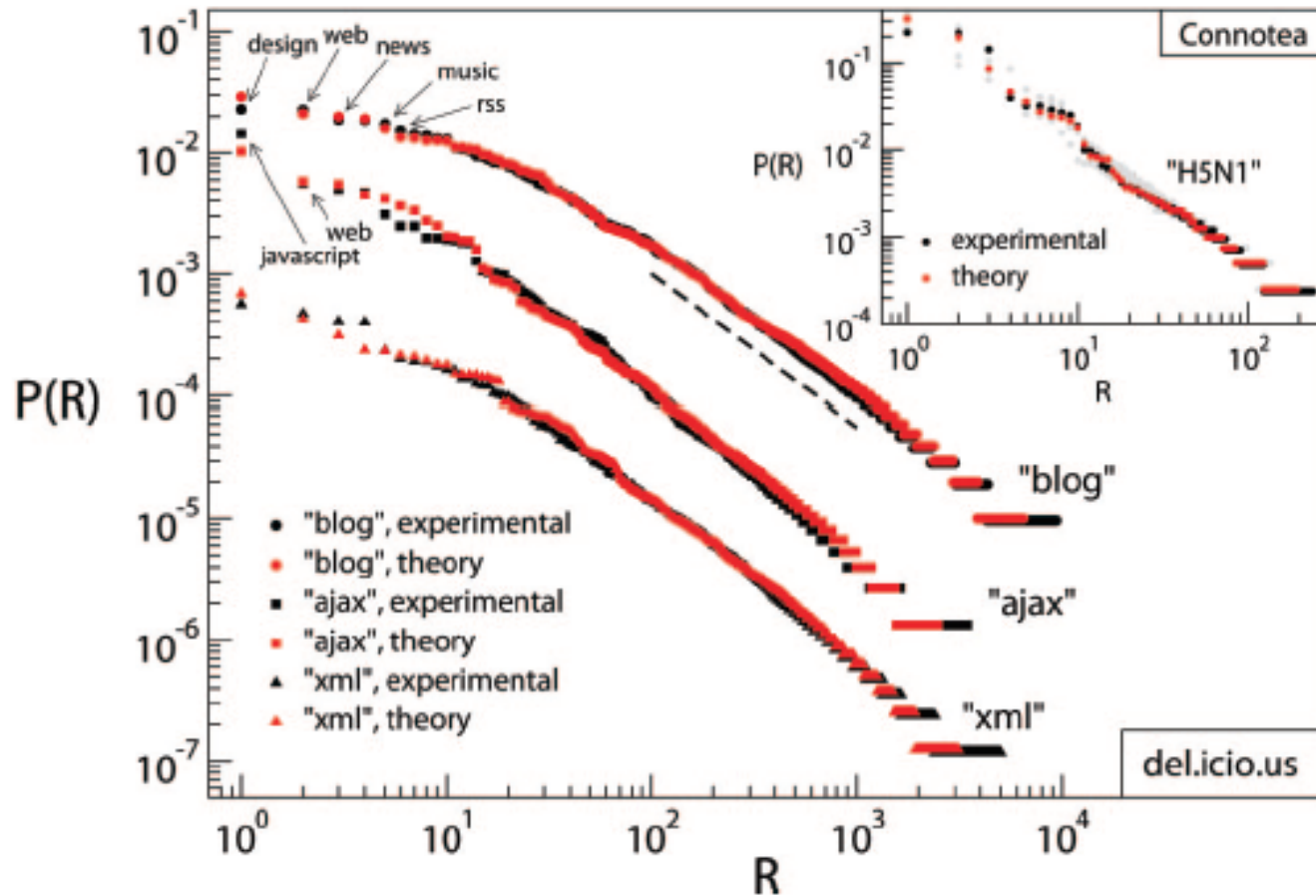
→ time

■ ajax

Source: Semiotic dynamics and collaborative tagging;

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1785269>

# tag popularity

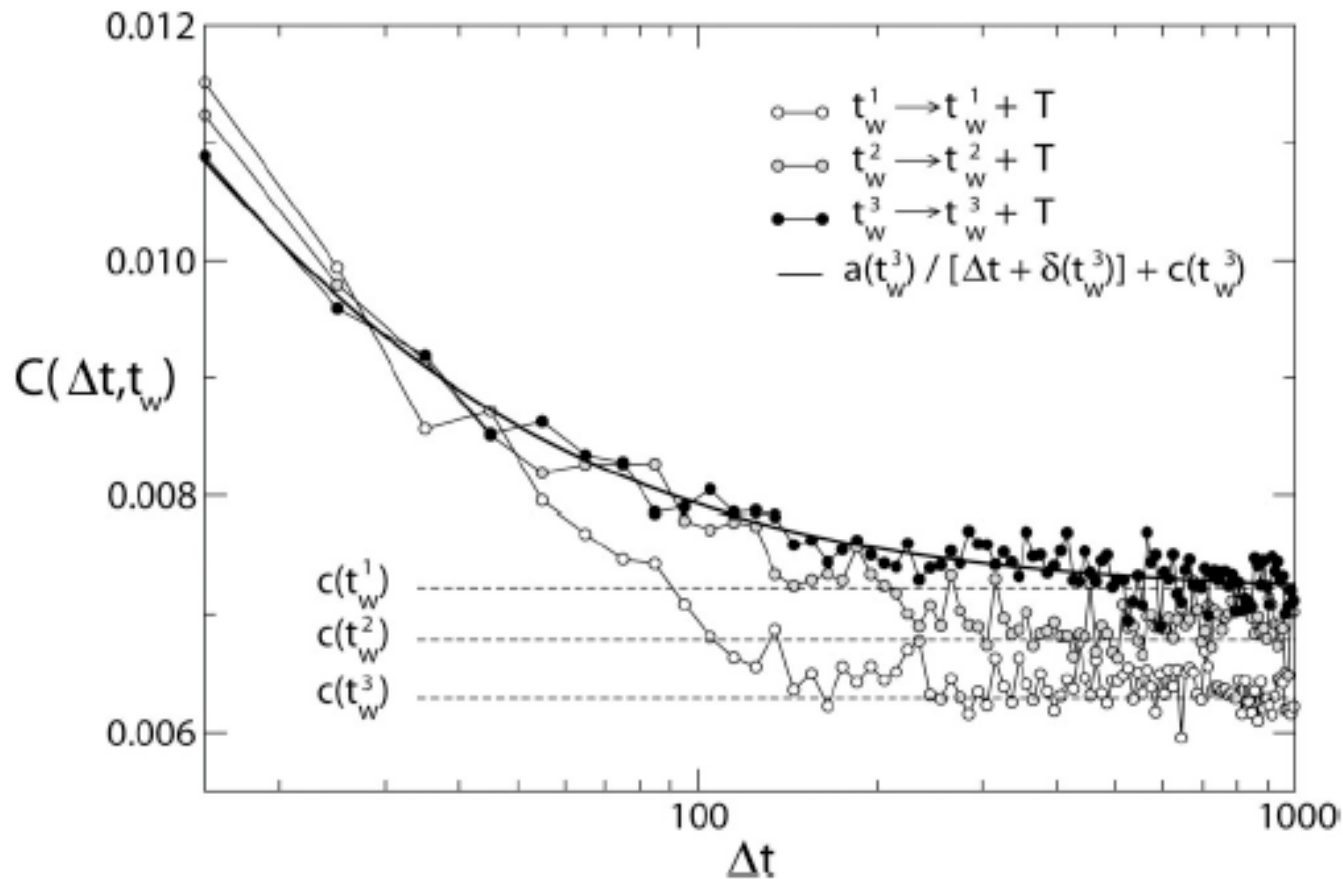


Source: Semiotic dynamics and collaborative tagging;

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1785269>

## will the same tag be used?

- as more time elapses, probability decays



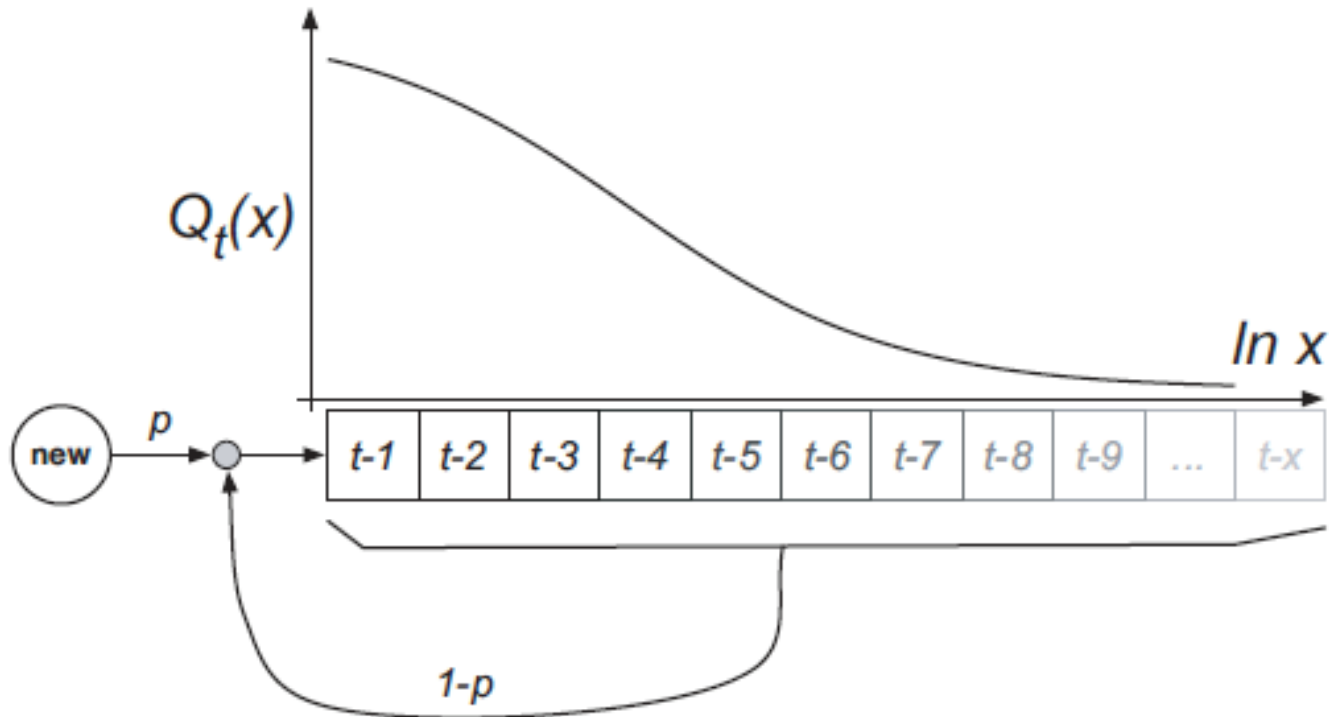
Source: Semiotic dynamics and collaborative tagging;

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1785269>



## tagging process

- Yule process: with probability  $p$ , choose new tag, with probability  $1-p$  copy an existing tag (but weigh by how long ago the tag was used...)



Source: Semiotic dynamics and collaborative tagging;

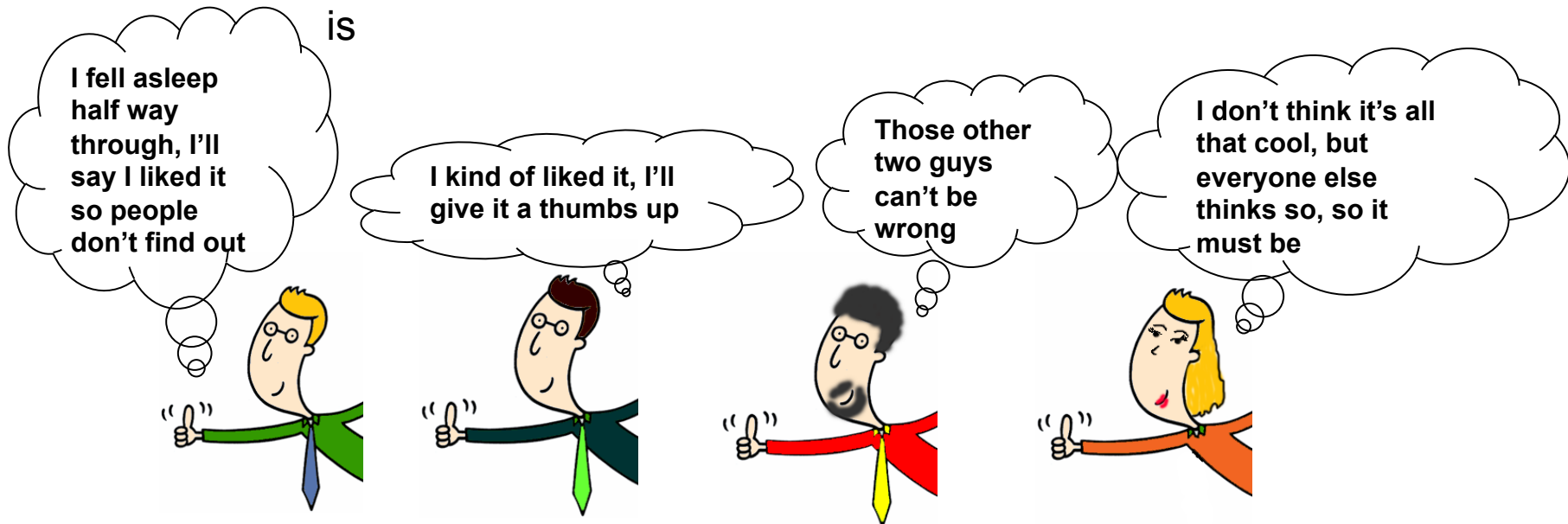
Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1785269>

# If collaborative filtering is so great, why do mediocre things sometimes become big hits, and true gems sometimes fall by the wayside?

## ■ Information cascades

### ■ herding behavior

- individual signal (knowledge, opinion)
- group signal (what others are saying)
- group can overpower individual signal
- things can become big hits, depending on what the word-of-mouth is



## Music Lab experiment at Columbia

 **MUSIC LAB**

 **COLUMBIA  
UNIVERSITY**



**FREE MUSIC DOWNLOADS**

Source: Music Lab, <http://www.musiclab.columbia.edu/>

## **Social influence study published in Science last week**

- **Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market**
  - **Matthew J. Salganik, Peter Sheridan Dodds, Duncan J. Watts**
- **Science, Feb. 10<sup>th</sup>, 2006**
- **Web experiment** <http://musiclab.columbia.edu/>
  - **set up site with free music downloads**
  - **14,000 participants (recruited through a teen-interest site)**
  - **profile information (age, gender, music influence, knowledge)**



[ [Security & Privacy](#) ] [ [Team](#) ] [ [Legal Issues](#) ] [ [FAQ](#) ] [ [Site Status](#) ] [ [Contact Us](#) ]

# WELCOME



Music Lab is a research project conducted by scientists from Columbia University to learn about how people form opinions about music. If you participate in Music Lab you will have a chance to download free new music.

After answering a few questions about yourself, you will be presented with a menu of songs by cool new artists. Your participation will take between 5 minutes and about two hours depending on how many songs you choose to listen to.

If you understand and are ready to participate, please click on the appropriate button below. If you would like to learn more about the research, please investigate the links at the top of the page. Enjoy.



How many people have chosen to download this song?

[Help] [Log off]

# of down loads

FORTHFADING: "fear"	13
SELSIUS: "stars of the city"	7
EMBER SKY: "this upcoming winter"	7
SALUTE THE DAWN: "i am error"	6
HARTSFIELD: "enough is enough"	6
STAR CLIMBER: "tell me"	5
DANTE: "lifes mystery"	4
BEERBONG: "father to son"	4
RYAN ESSMAKER: "detour_(be still)"	4
HALL OF FAME: "best mistakes"	4
THE FASTLANE: "til death do us part (i dont)"	3
STUNT MONKEY: "inside out"	3
THE CALEFACTION: "trapped in an orange peel"	3
SUM RANA: "the bolshevik boogie"	3
SILVERFOX: "gnaw"	3
PARKER THEORY: "she said"	3
SHIPWRECK UNION: "out of the woods"	3
NOT FOR SCHOLARS: "as seasons change"	3

00:30 sum rana - the bolshevik boogie

volume

|| ▶

Please rate this song.

You don't need to wait until it's finished

- ★ ★ ★ ★ ★ i love it
- ★ ★ ★ ★ i like it
- ★ ★ ★ it's OK
- ★ ★ i don't like it
- ★ i hate it

Rate It!

03:21 forthfading - fear

volume

|| ▶

Would you like to download this song?

Yes, download

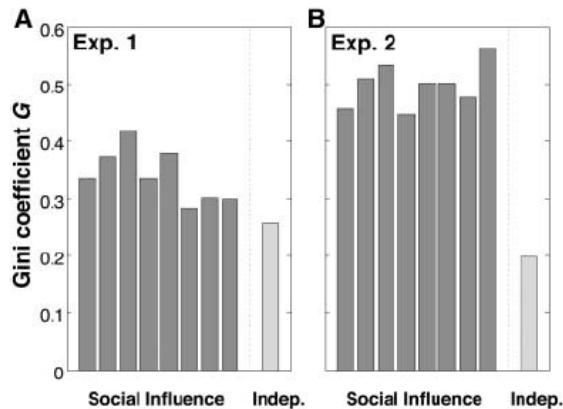
No, Thanks

## Experimental setup

- Subjects were randomly assigned to different groups
- 1 'independent' group: no information about downloads by others
- 8 'social influence groups'
  - see how many downloads were made by people in your own group (participants are unaware of the existence of groups, just of 'others')
  - Creates 8 different 'worlds' where the success or failure of a song evolves independently

## Findings about social influence

- Best songs rarely did poorly
- Worst songs rarely did well
- Anything else was possible!
- The greater the social influence, the more unequal and unpredictable the collective outcomes become.



- Experiment 2: songs shown in order of download popularity
- Experiment 1: songs shown in random order
- In both experiments variance in song success higher in the social influence case





## summary

- tagging networks are tripartite
- tagging is a process of invention and imitation
- imitation can skew popularity results