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
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...

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

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

<table>
<thead>
<tr>
<th>Paul</th>
<th>Tech Consulting</th>
<th>Systems Integration</th>
<th>32 docs viewed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>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.</td>
<td></td>
</tr>
</tbody>
</table>

**users similar to Paul Johansen**

<table>
<thead>
<tr>
<th>sim</th>
<th>name</th>
<th>unit</th>
<th>group</th>
<th>function</th>
<th>family</th>
<th>#docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.35</td>
<td>John</td>
<td></td>
<td></td>
<td>Solution Architect</td>
<td>Systems Integration</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>John</td>
<td></td>
<td></td>
<td>Solution Architect</td>
<td>Systems Integration</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Solution Architect</td>
<td>Systems Integration</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SEM</td>
<td>EMCI</td>
<td>Tech Consulting</td>
<td>Systems Integration</td>
<td>46</td>
</tr>
</tbody>
</table>

**Source:** Lada Adamic
Mapping knowledge communities from download patterns

Each node is a user accessing the system. Links identify users looking at the same documents. Color identifies position in the organization.

Users across the organization share interests based on the documents they access.
another example of expertise search, this time using occurrence of names in publicly available documents

PeopleFinder²

Search by:  
Person  
Department  
Topic

“A knowledge management” Group by Person

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

<table>
<thead>
<tr>
<th>Score</th>
<th>Name</th>
<th>Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.00</td>
<td>Bob</td>
<td>dept1</td>
</tr>
<tr>
<td>57.14</td>
<td>Sam</td>
<td>dept2</td>
</tr>
<tr>
<td>42.85</td>
<td>Kate</td>
<td>dept3</td>
</tr>
<tr>
<td>42.85</td>
<td>Harold</td>
<td>dept4</td>
</tr>
<tr>
<td>28.57</td>
<td>Richard</td>
<td>dept2</td>
</tr>
<tr>
<td>28.57</td>
<td>Uwe</td>
<td>dept5</td>
</tr>
</tbody>
</table>

See matches...
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")

<table>
<thead>
<tr>
<th>Score</th>
<th>Name</th>
<th>Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>100.00</td>
<td>Bob (dept1)</td>
<td></td>
</tr>
<tr>
<td>57.14</td>
<td>Sam (dept2)</td>
<td></td>
</tr>
<tr>
<td>42.85</td>
<td>Kate (dept3)</td>
<td></td>
</tr>
<tr>
<td>42.85</td>
<td>Harold (dept4)</td>
<td></td>
</tr>
</tbody>
</table>

- See matches...

<table>
<thead>
<tr>
<th>Score</th>
<th>Name</th>
<th>Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>Lada Adamic &gt; Victor &gt; Bob</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>Lada Adamic &gt; Victor &gt; Sam</td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>Lada Adamic &gt; Victor &gt; Bob &gt; Kate</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>Lada Adamic &gt; Victor &gt; Sam &gt; Kate</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>Lada Adamic &gt; Victor &gt; Bob &gt; Harold</td>
<td></td>
</tr>
</tbody>
</table>
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

**Finally: Official DS vs DS Lite comparison pictures! (including brightness)**
Submitted by DevilRejection 14 hours 8 minutes ago (via http://www.genmuy.com/showthread...)
No more concepts, no more CGI, someone who went to the DS Conference in Japan yesterday (Feb, 15, 2006)

**Students at MIT give flying car a shot**
Submitted by fuzzykit 15 hours 52 minutes ago (via http://theag.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

**iLife '06 Updates Available**
Submitted by snipetherick 15 hours 28 minutes ago (via http://www.macrumors.compage...)
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.

**Babies have an innate ability to do simple maths**
Submitted by Bunrasted 1 day 1 hour ago (via 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.

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

what’s this?

what is going on in the German blogosphere?

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

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 blogsphere...

So I am just adding Geekzone to the whole brrreeeport started by Scobble.

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

tag proportions – different tags for different people?

simple model of user behavior

- Poly'a'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

  ![Poly'a's urn diagram]

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

Source: del.icio.us, http://del.icio.us
tagging activity

- Catutto et al. PNAS 2006

time evolution

- blog
- ajax

Source: Semiotic dynamics and collaborative tagging;
tag popularity

will the same tag be used?

- as more time elapses, probability decays

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…)

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

I fell asleep half way through, I'll say I liked it so people don't find out

I kind of liked it, I'll give it a thumbs up

Those other two guys can't be wrong

I don't think it's all that cool, but everyone else thinks so, so it must be
Music Lab experiment at Columbia

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\(^{th}\), 2006**
- **Web experiment** [http://musiclab.columbia.edu/](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)
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.

<table>
<thead>
<tr>
<th>Artist</th>
<th>Title</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>FORTHFAVING</td>
<td>&quot;fear&quot;</td>
<td>13</td>
</tr>
<tr>
<td>SELSIUS</td>
<td>&quot;stars of the city&quot;</td>
<td>7</td>
</tr>
<tr>
<td>EMBER SKY</td>
<td>&quot;this upcoming winter&quot;</td>
<td>7</td>
</tr>
<tr>
<td>SALUTE THE DAWN</td>
<td>&quot;i am error&quot;</td>
<td>6</td>
</tr>
<tr>
<td>HARTSFIELD</td>
<td>&quot;enough is enough&quot;</td>
<td>6</td>
</tr>
<tr>
<td>STAR CLIMBER</td>
<td>&quot;tell me&quot;</td>
<td>5</td>
</tr>
<tr>
<td>DANTE</td>
<td>&quot;fires mystery&quot;</td>
<td>4</td>
</tr>
<tr>
<td>BEEBONGA</td>
<td>&quot;father to son&quot;</td>
<td>4</td>
</tr>
<tr>
<td>RYAN ESSMAKER</td>
<td>&quot;deelur (be still)&quot;</td>
<td>4</td>
</tr>
<tr>
<td>HALL OF FAME</td>
<td>&quot;best mistake&quot;</td>
<td>4</td>
</tr>
<tr>
<td>THE FASTLANE</td>
<td>&quot;till death do us part (i dont)&quot;</td>
<td>3</td>
</tr>
<tr>
<td>STUNT MONKEY</td>
<td>&quot;inside out&quot;</td>
<td>3</td>
</tr>
<tr>
<td>THE CALEFACTION</td>
<td>&quot;trapped in an orange peel&quot;</td>
<td>3</td>
</tr>
<tr>
<td>SUM RANA</td>
<td>&quot;the bolshevik boogie&quot;</td>
<td>3</td>
</tr>
<tr>
<td>SILVERFOX</td>
<td>&quot;gnaw&quot;</td>
<td>3</td>
</tr>
<tr>
<td>PARKER THEORY</td>
<td>&quot;she said&quot;</td>
<td>3</td>
</tr>
<tr>
<td>SHIPWRECK UNION</td>
<td>&quot;out of the woods&quot;</td>
<td>3</td>
</tr>
<tr>
<td>FOR NOT SCHOLARS</td>
<td>&quot;as seasons change&quot;</td>
<td>3</td>
</tr>
</tbody>
</table>

How many people have chosen to download this song?

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
In the lab:

- Analyze a tagging network of I-Schools based on del.icio.us data
summary

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