University of Arkansas at Little Rock

Understanding Social Media: Tools, Applications, and Processes

Nitin Agarwal, Information Science, UALR

nxagarwal@ualr.edu

http://ualr.edu/nxagarwal/sbp11_tutorial.pdf

This work is supported in part by the grants from the US Office of Naval Research (ONR).


barrier
beliefs
advance
express
enables
discussed
mining
individual
observed
offer
objective
overview
plethora
still
provide
remain
users
wisdom
well
vast
storehouse
Tunisian
behavioral
concepts
Egyptian
dynamic
discovered
Nonetheless
manifested
socio-technical
standards
symbiosis
transformed
community
diversity
graph-theoretic
processes
open-source
platform
share
unorganized
applications
capabilities
exploited
induced
modeling
intelligence
participate
publication
created
computational
inexpensive
media
privacy
science
social
amounts
highlighting
phenomenon
issues
opportunities
Techniques
comprehensive
environment
organizations
changes
collective
new
methodologies
ranging
research
understanding
transnational
extraction
organic
overwhelming
approaches
easy-to-use
consumers
management
low
recent
scales
challenges
demonstrated
exemplified
studies
knowledge
voice
socio-political
clustering
collaborative
discussions
Specifically
creating
emergent
information
un-regulated
acts
collection
insights
movements
producers
democratic
members
evaluation
synergies
ubiquitous
almost
content
disaster
influence
opinions
thoughts
suitable
case
crisis
give
model
Internet
opinion
networks
open
recovery
trust
tutorial
data
various
uprisings
way

Outline

- Social Media
  - Characteristics
  - Taxonomy
- Data collection
  - APIs
  - Available datasets
- Data analysis - techniques/algorithms
  - Graph theoretic
  - Content based
- Applications, Current Research Trends, & Opportunities
  - Influence, familiar strangers, crisis response, collective action

SOCIAL MEDIA & WEB 2.0

Social Media

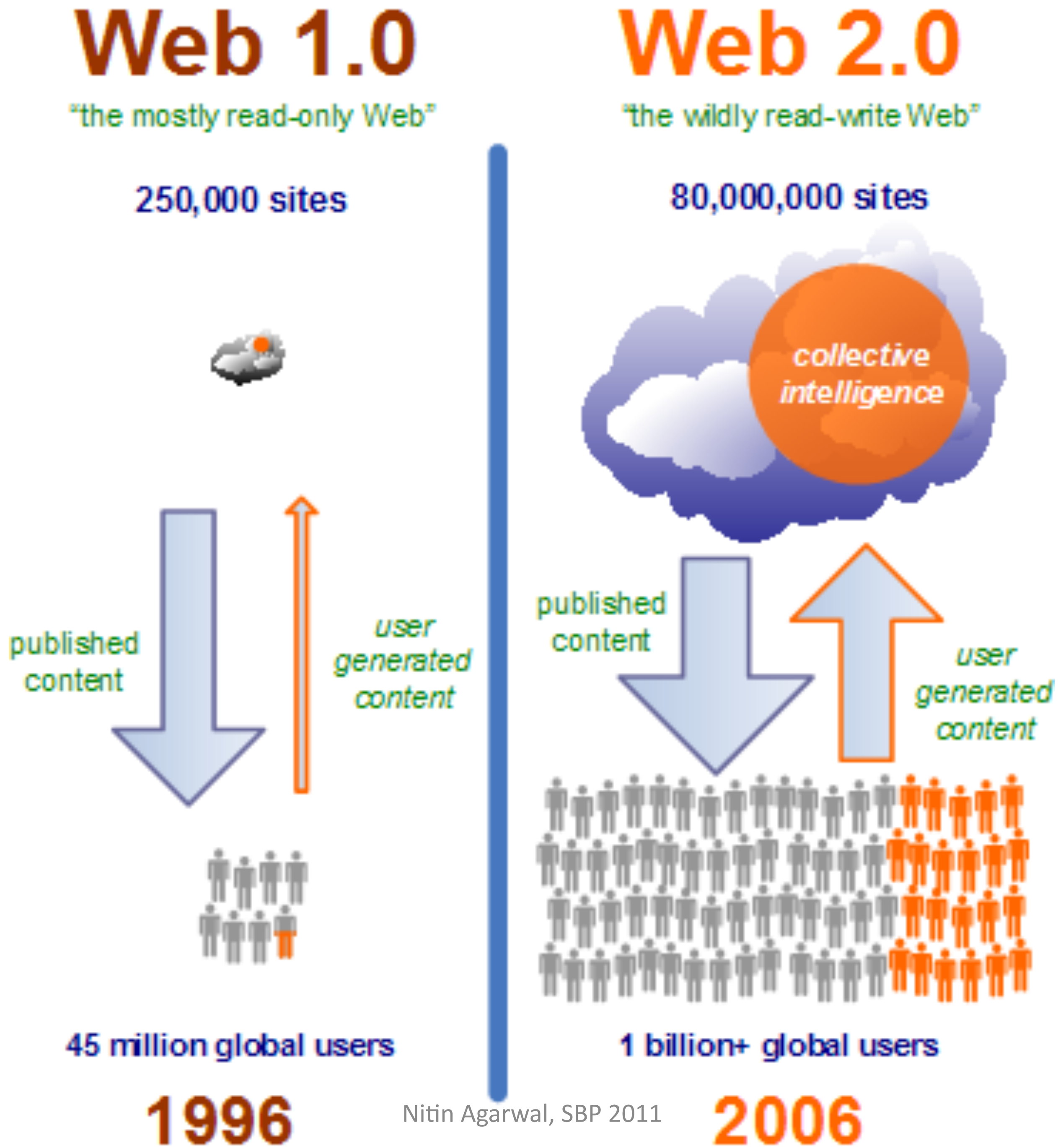
- Media designed to be
  - disseminated through social interactions,
  - created using highly accessible and scalable publishing techniques,
  - using Internet and web-based technologies to transform monologues (one to many) into social media dialogues (many to many)
  - It supports the democratization of knowledge and information, transforming people from content consumers to content producers.
- User-generated content (UGC)/Consumer-generated media (CGM)
- Web 2.0

Web 2.0

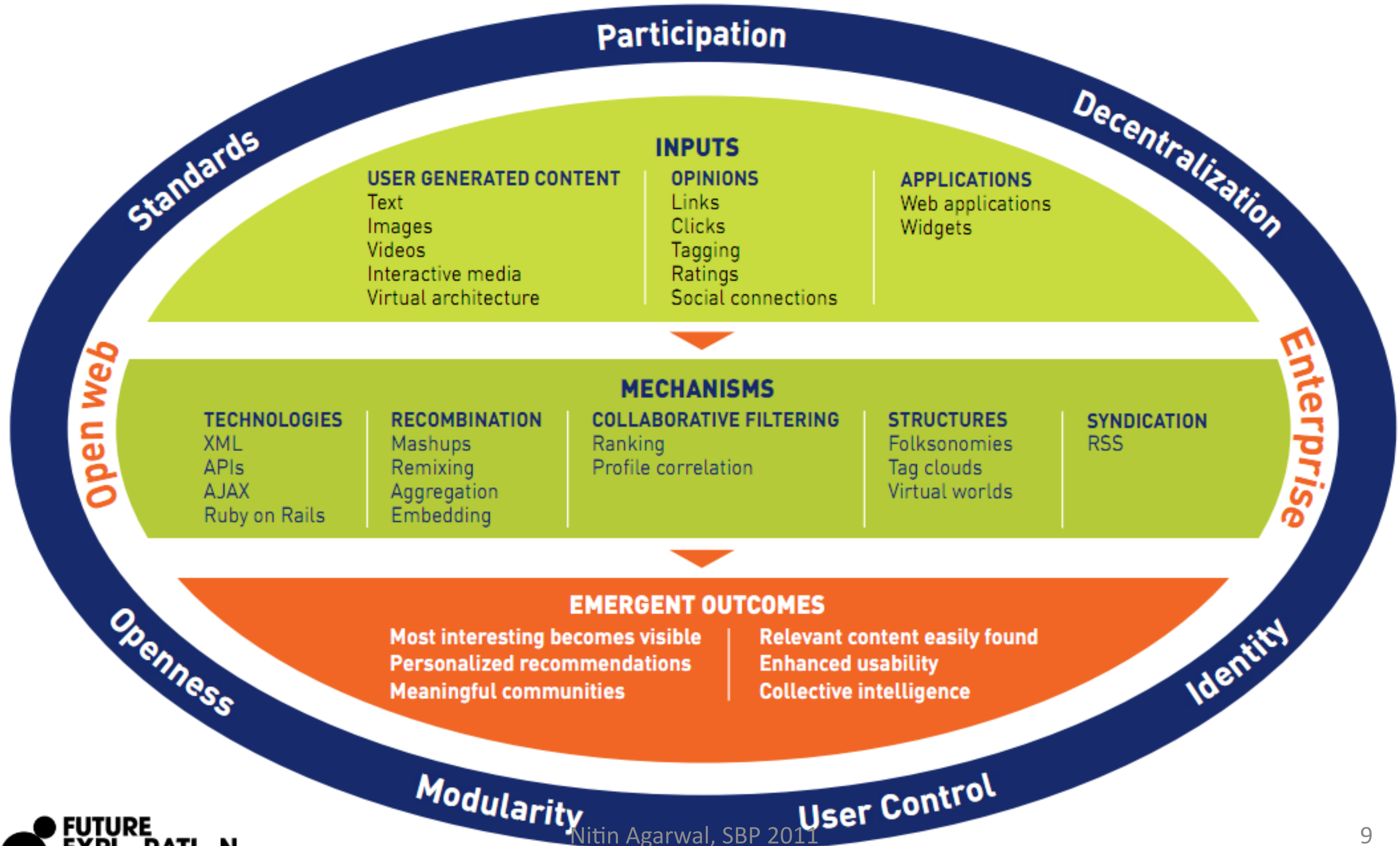
- Introduced in 1999, popularized in 2004
- Tim Berners-Lee called it the Read/Write Web
- Rather a paradigm shift than a technology shift
- Web as platform
- Different perspectives
  - "Social web", "participatory web"
  - "Standardized web
- Openness, freedom, collective intelligence
- Customers are building your business

Web 2.0


DECEMBER 25, 2006 / JANUARY 1, 2007
www.time.com
TIME
PERSON OF THE YEAR
You.
00:00 / 20:06
Yes, you.
You control the Information Age.
Welcome to your world.


Web 1.0
"the mostly read-only Web"
250,000 sites
published content
user generated content
45 million global users
1996
Web 2.0
"the wildly read-write Web"
80,000,000 sites
collective intelligence
published content
user generated content
1 billion+ global users
2006

Web 2.0 Architecture



FUTURE EXPLORATION NETWORK

Web 2.0 Architecture

- Participation – easy content creation and sharing by anyone
- Standards – content retrieval and integration
- Decentralization – from content creation to content storage
- Openness – open and transparent access to content and applications
- Modularity – highly component-oriented development
- User control – content, activities, and identity
- Identity – reveal/hide upon user's discretion

Future of Web 2.0 & Social Media

- What do you think?
- Lets find out…

Social Media Characteristics

- Industrial media
  - Traditional, broadcast, or mass media
- Factors distinguishing Industrial/Social Media
  - **Accessibility**: available to anyone
  - **Permanence**: dynamic
  - **Reach**: global audience
  - **Recency**: interactive and responsive
  - **Usability**: almost zero operational costs

Social Media - Categories

Category	Social Media Sites
Social Signalling	Blogs (Wordpres, Blogger), Microblogs (Twitter), Friendship networks (Facebook, MySpace, LinkedIn, Orkut)
Social Bookmarking	Del.icio.us, StumbleUpon
Media Sharing	Flickr, Photobucket, Youtube, Megavideo, Justin.tv, Ustream
Social News	Digg, Reddit
Social Health	PatientsLikeMe, DailyStrength, CureTogether
Social Collaboration	Wikipedia, Wikiversity, Scholarpedia, AskDrWiki
Social Games	FourSquare, FarmVille, SecondLife, EverQuest (Virtual worlds)
Q & A	Yahoo Answers,Quora

Top 20 Most Visited Websites

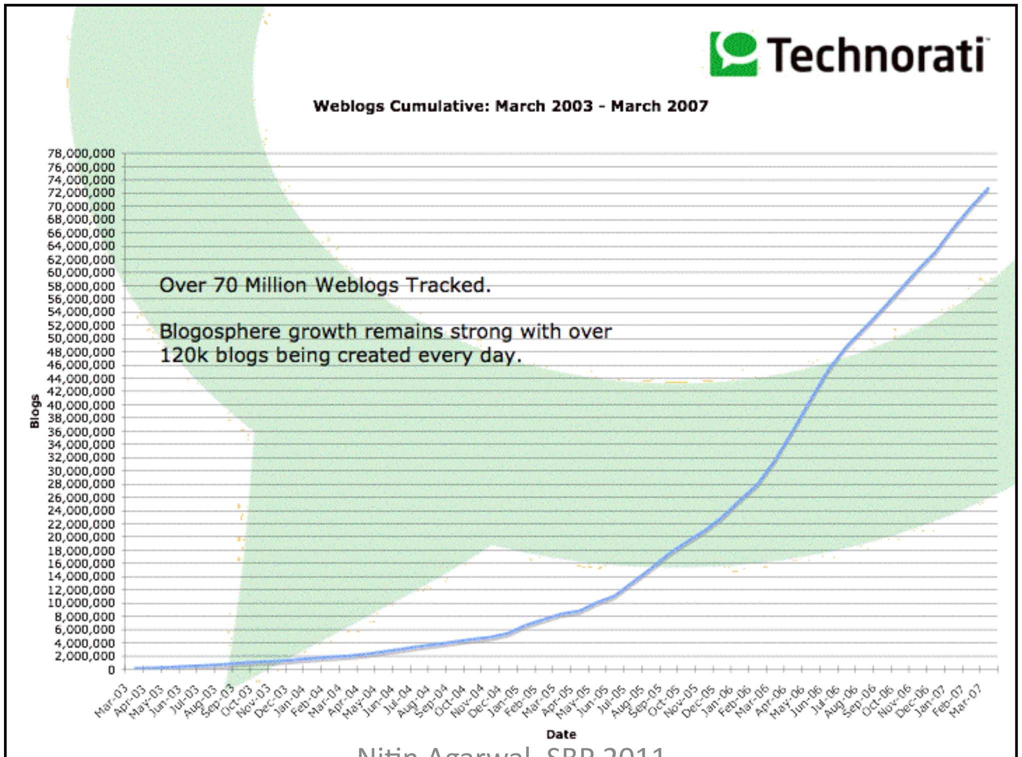
- Internet traffic report by Alexa on January 19, 2011

1	Google	11	MSN
2	**Facebook**	12	Yahoo! Japan
3	**YouTube**	13	Taobao.com
4	Yahoo!	14	Amazon
5	Windows Live	15	Google India
6	Baidu	16	Sina.com.cn
7	**Blogger**	17	Google Germany
8	**Wikipedia**	18	Google Hongkong
9	QQ	**19**	**Wordpress**
10	**Twitter**	20	Bing

- 50% of the top 10 websites are social media sites.

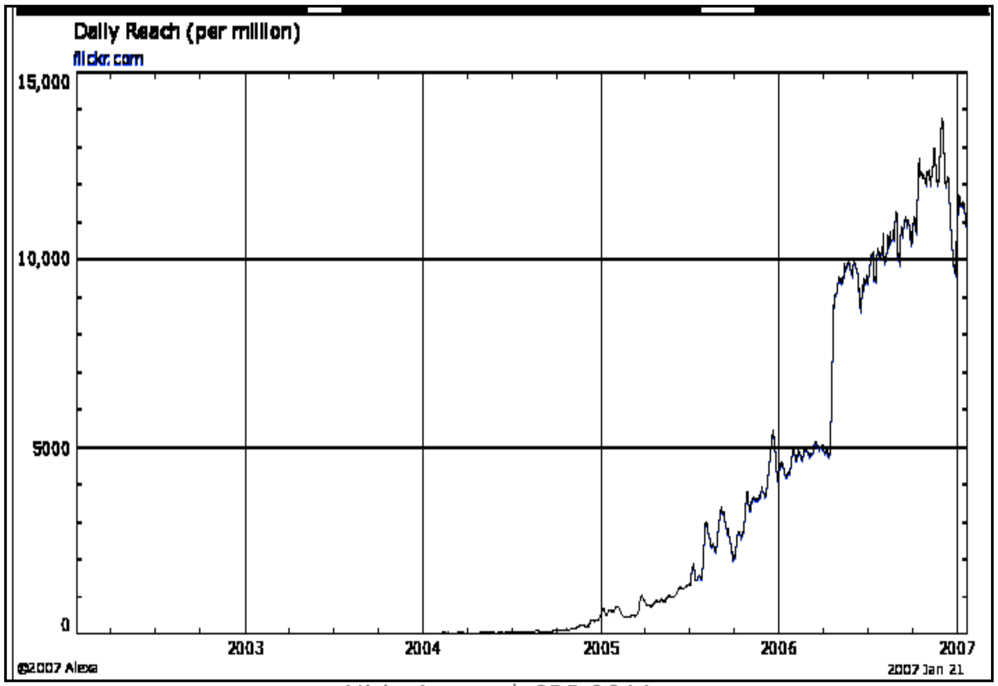
Blogosphere Growth

- January 2011: Blogpulse indexed over 153 million blogs
- 80,731 new blogs per day = 1 new blog per second
- 1,186,637 = 13.73 new blog posts per second



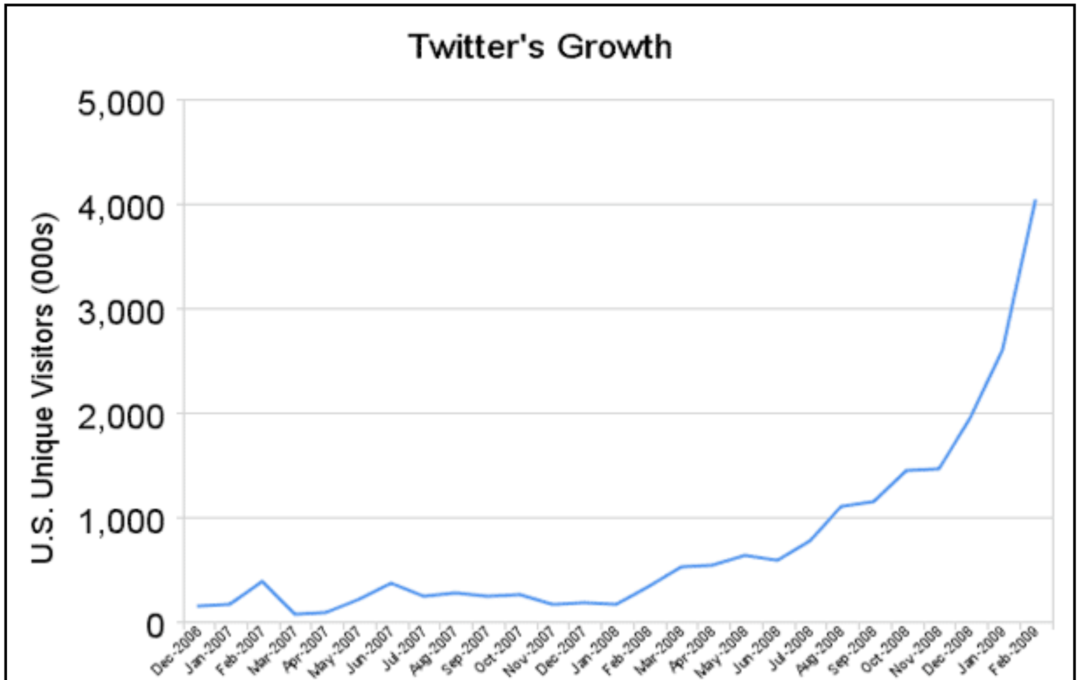
Flickr Growth

- Over 10 million users - as of June 2009
- 242% annual growth rate [Mislove *et al.* 2008]
- 3.6 billion images and tags



Twitter Growth

- 19.5 million users - March 2009
- 1382% annual growth
- 3 million tweets per day (34.7 tweets per second)



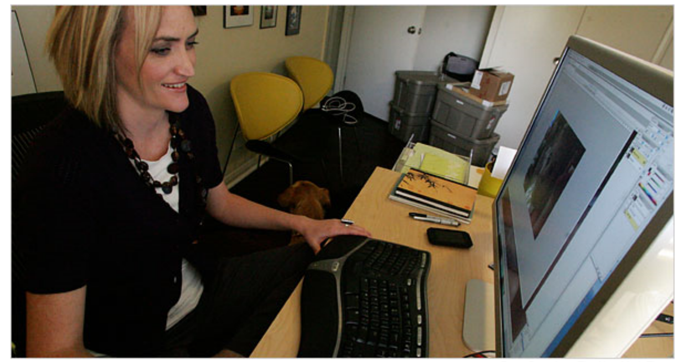
Blogs – Impacts and Value

My Son, the Blogger: An M.D. Trades Medicine for Apple Rumors

Arnold Kim, founder and senior editor of MacRumors.com.

"The site places MacRumors No. 2 on a list of the '25 most valuable blogs,' …" What is the potential value? "Two of the other tech-oriented blogs on its list, …, were sold earlier this year, reportedly for sums in excess of $25 million."

Woman to Woman, Online



Bess Greenberg/The New York Times

"The site, chock full of advertising, is a moneymaking machine – so much so that Ms. Armstrong and her husband have both quit their regular jobs."
The reason? The advertisers are eager to influence her 850,000 readers.

"Queen of the Mommy Bloggers"

The New York Times

Source: The New York Times

Blogs – Impacts and Value

Harnessing the Power of the Mom Blogger

The mother bloggers can become "ambassadors of brands," said Sarah Hofstetter, senior vice president for emerging media and brand strategy at 360i, a digital agency owned by Dentsu, the Japanese advertising agency. "These mom bloggers have tremendous personality and tremendous opinions."

Businesses and Twitter

The New York Times

Curtis Kimball, owner of a crème brûlée cart in San Francisco, uses Twitter to drive his customers to his changing location.

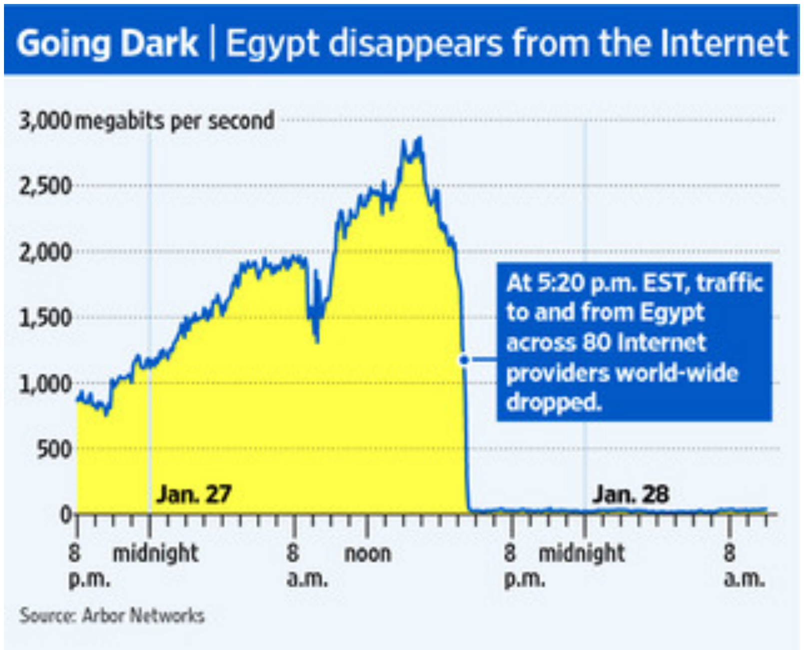
Source: http://www.nytimes.com/2009/07/23/business/smallbusiness/23twitter.html

paidContent.org
THE ECONOMICS OF DIGITAL CONTENT

Hedge Fund Is Betting That Twitter Is Wall Street's Crystal Ball

Source: http://bit.ly/glH4mv

Socio-Political Dynamics and Twitter



http://twitter.com/speak2tweet →

Challenges

- Time Challenge: Dynamic environment
  - Data gets stale too soon
- Size Challenge: Phenomenal growth
  - Difficult to follow
- Sparse link structure
  - Often do not cite the source
- Information Quality
  - Colloquial, slang text, e.g., “Arrghhh!!” “coooooool”
  - Lots of off-topic chatter/noise
  - Intentional misspellings
  - Abbreviations, cryptic texts, smileys { :) :(} “Twitter vocabulary”

Challenges (contd.)

- Spam
  - Nearly 45% of conversation on Twitter is babble
  - How Much Can A Spammer Pocket A Day? You'd Be Surprised – NPR (http://n.pr/hZi45C)
- Privacy and Security of personal information
- Dangers of inaccurate information
  - Relevance vs. Reliability

Linked in

Andi Fisher

Senior Manager, Global Internet Marketing at Dolby Laboratories, Inc. at Dolby

San Francisco Bay Area

Current
- **Chief Go To Gal at Your Online Go To Gal LLC**
- **Senior Manager, Global Internet Marketing at Dolby Laboratories, Inc. at Dolby**

Past
- Online Marketing Manager, Internet Marketing at Logitech
- Global Program Manager, Internet Marketing at Logitech
- Global Program Manager, Internet Development at Logitech

9 more...

Recommended **12** people have recommended Andi

Connections **361** connections

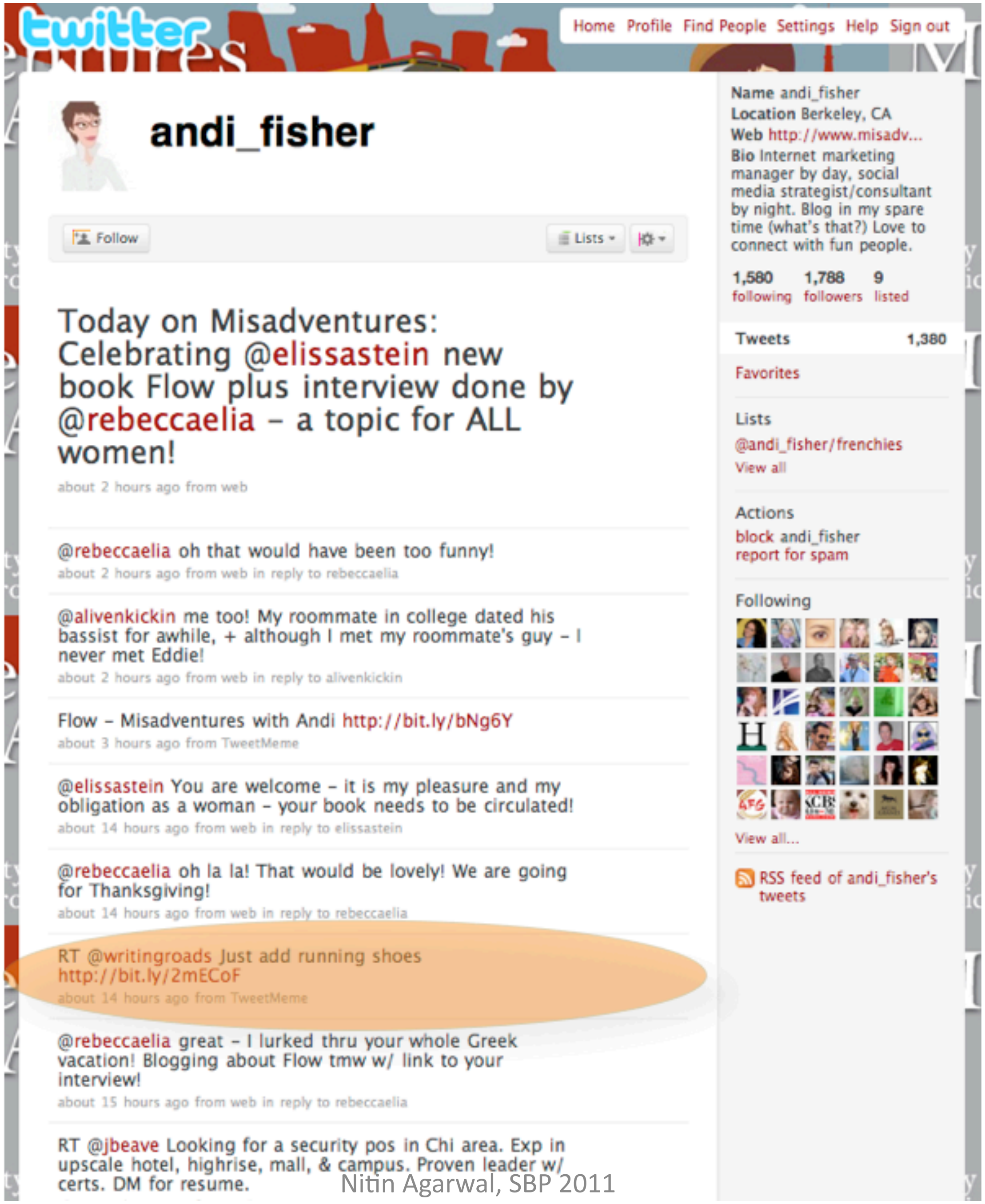
Industry Marketing and Advertising

Andi Fisher's Summary

- Seasoned leader with over 10+ years experience impacting the performance of companies through successfully launching websites and developing programs that capitalize on the company's online objectives.
- Social media strategist and enabler working with small businesses to help them to determine their social media needs.
- Design project management procedures and evangelize content management systems to deliver planned goals.
- Global Program Manager focused on improving product/team performance
- Successfully directed numerous projects globally.
- Capable of gracefully navigating across multiple concurrent projects.
- Develop end-to-end project management processes and communication methodologies.
- Extensive experience in training people on systems and processes.
- Exceptional interpersonal skills.

Andi Fisher's Specialties:

Online Marketing, Global Web Program/Project Management, Localization, HTML, Usability, International Experience, Training, Social Media, Blogging


twitter
Home Profile Find People Settings Help Sign out
andi_fisher
Follow
Lists
Today on Misadventures: Celebrating @elissastein new book Flow plus interview done by @rebeccaelia – a topic for ALL women!
about 2 hours ago from web
@rebeccaelia oh that would have been too funny!
about 2 hours ago from web in reply to rebeccaelia
@alivenkickin me too! My roommate in college dated his bassist for awhile, + although I met my roommate's guy – I never met Eddie!
about 2 hours ago from web in reply to alivenkickin
Flow – Misadventures with Andi http://bit.ly/bNg6Y
about 3 hours ago from TweetMeme
@elissastein You are welcome – it is my pleasure and my obligation as a woman – your book needs to be circulated!
about 14 hours ago from web in reply to elissastein
@rebeccaelia oh la la! That would be lovely! We are going for Thanksgiving!
about 14 hours ago from web in reply to rebeccaelia
RT @writingroads Just add running shoes http://bit.ly/2mECoF
about 14 hours ago from TweetMeme
@rebeccaelia great – I lurked thru your whole Greek vacation! Blogging about Flow tmw w/ link to your interview!
about 15 hours ago from web in reply to rebeccaelia
RT @jbeave Looking for a security pos in Chi area. Exp in upscale hotel, highrise, mall, & campus. Proven leader w/ certs. DM for resume.
Name andi_fisher
Location Berkeley, CA
Web http://www.misadv...
Bio Internet marketing manager by day, social media strategist/consultant by night. Blog in my spare time (what's that?) Love to connect with fun people.
1,580 following
1,788 followers
9 listed
Tweets 1,380
Favorites
Lists
@andi_fisher/frenchies
View all
Actions
block andi_fisher
report for spam
Following
View all...
RSS feed of andi_fisher's tweets


blogcatalog
explore the blogosphere.
home
directory
search
discuss
groups
Blog Detail
Misadventures with Andi
http://www.misadventureswithandi.com
Merry musings of a feisty foodie slash lit-chickie slash globe-trotting wannabe Frenchie!
Added on: Jan 20th, 2009
Country: United States
Language: English
Author: andi Fisher
Listed in: Personal
Tags: books life movies food french

Google: A New Tool For U.S. Intelligence?

Source: http://n.pr/gDxzpR

“The traditional intelligence community is absolutely biased toward classified information,” said Lt. Col. Reid Sawyer, an Army intelligence officer and head of West Point's Combating Terrorism Center. "*I think that open source provides a critical lens into understanding the world around us in a much more dynamic way than traditional intelligence sources can provide*.”

Open sources include newspapers, local radio shows and, of course, Facebook and Twitter. The problem, intelligence officials will tell you, is tapping into all of that in a systematic way.

Please Rob Me: http://pleaserobme.com/

Information Relevance vs. Reliability

How to Prevent Restless Leg Syndrome:

An example

Search

NIH

Restless Legs Syndrome Fact Sheet: National Institute of ...
Restless legs syndrome (RLS) is a neurological disorder characterized by keep their legs in motion to minimize or **prevent** the sensations. ...
What is restless legs syndrome? - What are common signs and ...
www.ninds.nih.gov/.../**restless_legs**/detail_**restless_legs**.htm - Cached - Si

Wiki

How to Prevent Restless Leg Syndrome (RLS) - wikiHow
May 14, 2009 ... wikiHow article about **How to Prevent Restless Leg Synd**
How to Avoid Feet and Leg Problems if Standing for Work · How to Use a ..
www.wikihow.com/**Prevent-Restless-Leg-Syndrome**-(RLS) - Cached - Simi

eHow

How to Prevent Restless Leg Syndrome | eHow.com
How to Prevent Restless Leg Syndrome. The prevention of restless leg sy more an effort to eliminate or control other conditions which ...
www.ehow.com › ... › Restless Leg Syndrome - Cached - Similar

Health portal

How To Prevent Restless Leg Syndrome - HealthCentral
How To Prevent Restless Leg Syndrome .... Restless leg syndrome (RLS your waist size, a new study has found. A study of 88000 adults found ...
www.healthcentral.com/.../**how-to-prevent-restless-leg-syndrome**.html - C

News website

Restless Leg Syndrome Treatment - Restless Leg Syndror
Oct 12, 2009 ... Multimedia. Patient Voices: **Restless Leg Syndrome** Inter
Quinine had been widely used to **prevent** leg cramping. ...
health.nytimes.com › ... › r › Restless Leg Syndrome - Similar

Helpguide

Restless Legs Syndrome (RLS): Symptoms, Treatment, and
Exercises for **Restless Leg Syndrome** – Offers suggestions for types of e physical activity to include in your weekly routine to **prevent** restless ...
www.helpguide.org/life/**restless_leg_syndrome**_rls.htm - Cached

Top 6 results on Google (Feb 2, 2011)

Search string: How to prevent restless leg syndrome

Second Hit: WikiHow

1 **One of the best ways to rid yourself of RLS is to decrese your consumption of orange juice.** It is not currently known why the frequent intake of this breakfast beverage causes problems in some people. The cause could very well lie in the fact that most orange juice sold in the United States is imported from countries which do not adhere to stirct standards regarding the use of pesticides. When these imported fruits are used, residual qauntities of pesticides may be contaminating the juice and causing allergic reactions in some individuals. If you consume large amounts of orange juice, stop drinking it for several days. This may well releive your symptoms.

- Collaborative editing site
- Article claims that drinking orange juice causes RLS
- Dubious?

Is it really the case?

- First hit: NIH fact sheet
- Third hit: eHow
- Fourth hit: HealthCentral (managed by health experts)
  - No mention of orange juice but all say that iron deficiency is a possible reason. This is never mentioned on wikiHow.
- Fifth hit: New York Times
  - Mentions that orange juice could reduce RLS

Blogs

- A sample blog, blog post, tags, comments, archive, blog roll, ping, trackback, etc.
- Individual / Community blogs
- Anonymous / non-anonymous comments
- Blogcatalog - metadata

Wikis

- Wikipedia
  - Collaborative encyclopedia
  - Edit history (log) rollback
  - Watchlist
  - Discussion

Microblogging

- Twitter
  - Tweets (140 chars)
    - Links
  - Following
  - Followers
  - #tags
- Tweet Statistics

Media Sharing

- Flickr
  - Upload images
  - Tags
  - Contacts (friends)
  - Public groups (communities)

Social Bookmarking

- Delicious/Diigo/StumbleUpon
  - Tags
  - Network
- Fresh bookmarks
- Popular bookmarks
- Tags/folksonomy

Social News

- Digg
  - Ratings for stories
  - Friends
  - Recommendations
  - Popular (most diggs) / upcoming (most recent) stories / influential stories

Collaborative Answering

- [Yahoo! Answers]
  - Recent / Popular questions
  - Rate questions/rate answers
  - Question categories
  - User profile/network/fans
  - User scores

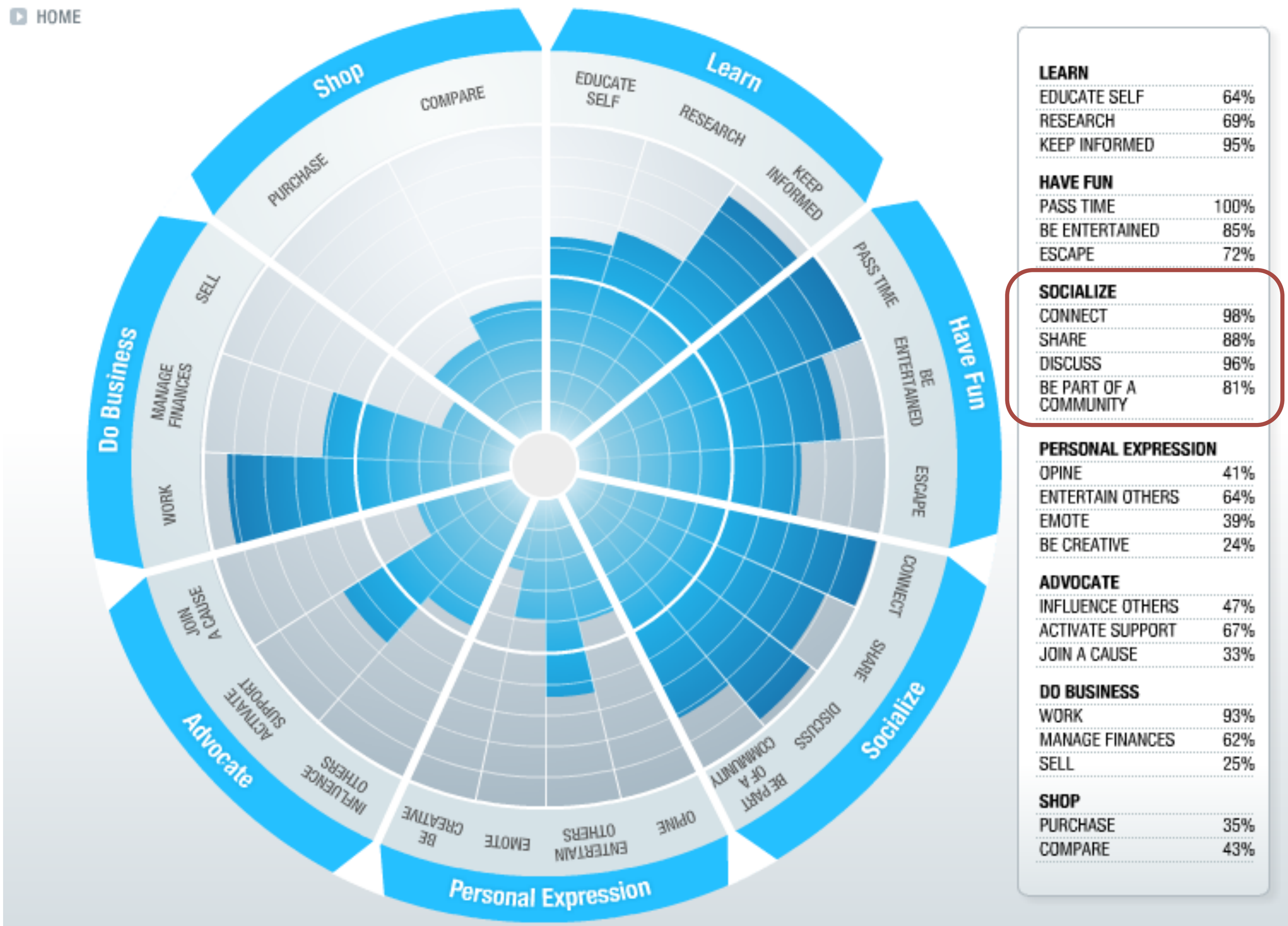
Friend Networks

- Facebook/Myspace
- LinkedIn
  - Professional networks
  - Hiring and recruitment

Social Media Aggregators

- Mixx
  - Combines feeds from Facebook, Digg and Twitter
  - Aggregate user/customer's reactions/interactions scattered across the landscape of social media on the publishers content
  - Promotes engagement and recirculation
- Surphace
  - Digs up old yet relevant content and links to the new content.
  - Engadget using Surphace
- YackTrack
  - Conversation tracker
  - "Yackability" measures how much conversation is occurring for a particular search.
- ConvoTrack
- Increase time-on-site, PVs per visit, CTRs for online ads

Social Applications on Mobile Devices



Source: http://www.intentindex.com/mobile/

Social Applications on Mobile Devices

- A recent study by Ruder Finn pointed out
  - 91% of the mobile subscribers engage in social computing applications as compared to the 79% of the desktop users.
  - People in the US on average spend 2.7 hours per day on mobile devices, of which
    - 45% post comments on social networking sites,
    - 43% connect with friends on social networking sites,
    - 40% share content with others, and
    - 38% share photos, making it a favorable platform for socializing.

DATA COLLECTION

Data Crawling

- API
- Webpage scraping
  - Nutch - http://wiki.apache.org/nutch/NutchTutorial
  - Open source topical crawlers, http://informatics.indiana.edu/fil/IS/JavaCrawlers/
  - Heretrix: http://crawler.archive.org/
- Blog Archive
  - Regular Expressions
- RSS feeds
  - Most Recent blogs
  - XML parsing *(also in APIs)*
  - Well defined structure
  - Feed aggregators, e.g., Feed on Feeds (http://feedonfeeds.com/)

API

- Application Programming Interface
- HTTP request
- Session tracking through API keys
  - Impose limits on usage
  - Tracks who is using
- Formats: XML, JSON
  - Interoperability
- Query parameters
  - API key
  - Other parameters (specific to API query)

Available APIs

- BlogCatalog (blog site details, blogger details)
- Twitter (friends, followers, tweets, etc.)
- Delicious (bookmarks, community tags, etc.)
- Technorati (blog site details, inlinks, etc.)
- Digg (popular stories, fresh stories, etc.)
- Facebook (graph)
- … and many more (http://www.programmableweb.com/)

BlogCatalog API

http://www.blogcatalog.com/api/

- `getinfo` query

API url: **`http://api.blogcatalog.com/getinfo?bcwsid=[apikey]&username=johndoe`**

Sample Response

```
<result>
 <user id="56848">johndoe</user>
 <realname>John Doe</realname>
 :
 <weblogs>
   <weblog id="4286052">
     <name>The JohnDoe Blog</name>
     <url>http://blog.johndoe.com</url>
      <bcurl>http://www.blogcatalog.com/blogs/the-johndoe-blog.html
     </bcurl>
     :
   </weblog>
 </weblogs>
```

BlogCatalog API

- `bloginfo` query

API url: **`http://api.blogcatalog.com/bloginfo?bcwsid=[apikey]&url=http://www.john-doe-blog.com`**

Sample Response

```
<categories>
    <category>Blog Resources</category>
    <category>Blogging</category>
</categories>
<tags>
    <tag>annoucements</tag>
    <tag>blogcatalog</tag>
    <tag>news</tag>
</tags>
```

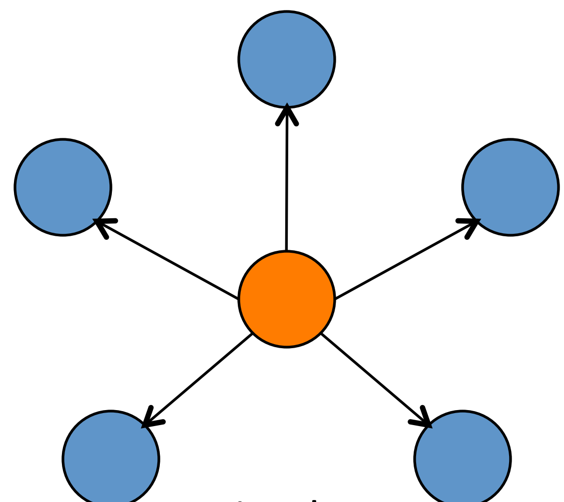
Twitter API

- http://dev.twitter.com/doc/
- No API key, 150 requests per hour
- `friends/ids` query

API url: http://api.twitter.com/1/friends/ids.xml?user_id=18872235

User ids


```xml
<?xml version="1.0" encoding="UTF-8"?>
<ids>
  <id>1401881</id>
  <id>6761692</id>
  <id>6636732</id>
  <id>813286</id>
  <id>7057722</id>
  :
</ids>
```

Friends

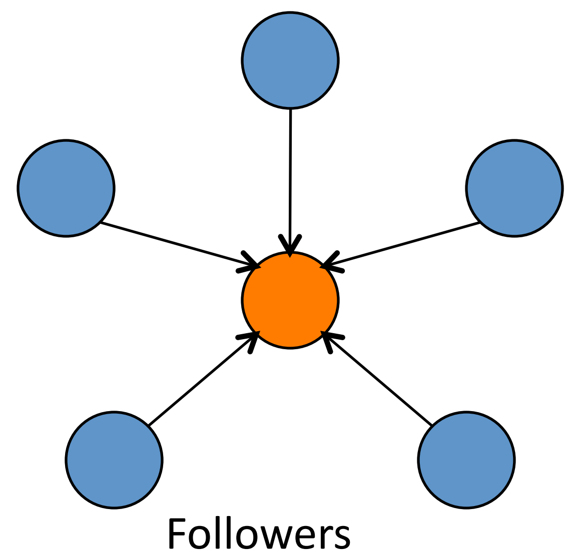
Twitter API

- `followers/ids` query

API url: http://api.twitter.com/1/followers/ids.xml?screen_name=buzzbissinger

Sample response:


```
<?xml version="1.0" encoding="UTF-8"?>
<ids>
 <id>683643</id>
 <id>744883</id>
 <id>755002</id>
 <id>611823</id>
 ⋮
</ids>
```

Twitter API

- `users/show` query – returns extended information of a given user.

API url: http://api.twitter.com/1/users/show.xml?user_id=18872235

Sample response:


```
– <user>
    <id>18872235</id>
    <name>buzzbissinger</name>
    <screen_name>buzzbissinger</screen_name>
    <location>Philadelphia</location>
  – <description>
      Author of Friday Night Lights, Prayer for City, 3 Nights in August. Cont. editor Vanity Fair
    </description>
  – <profile_image_url>
      http://a1.twimg.com/sticky/default_profile_images/default_profile_4_normal.png
    </profile_image_url>
    <url>http://buzzbissinger.com</url>
```

Twitter API

- http://dev.twitter.com/doc/get/trends
- Shows trending topics
  - Current: http://api.twitter.com/1/trends/current.json
  - Daily: http://api.twitter.com/1/trends/daily.json
  - Weekly: http://api.twitter.com/1/trends/weekly.json
- Response only in JSON format

Delicious API

API: http://www.delicious.com/help/api

Returns a list of tags and number of times used

https://api.del.icio.us/v1/tags/get

Requires authentication

Sample response

```
<tags>
  <tag count="1" tag="activedesktop" />
  <tag count="1" tag="business" />
  <tag count="3" tag="radio" />
  <tag count="5" tag="xml" />
  <tag count="1" tag="xp" />
  <tag count="1" tag="xpi" />
</tags>
```

Technorati API

- API can give additional information
- How do you track inlinks?
- `bloginfo` query

API URL: **`http://api.technorati.com/bloginfo?key=[apikey]&url=[blog url]`**

Sample response:

```
<result>
  <url>[URL]</url>
  <weblog>
    <name>[blog name]</name>
    <url>[blog URL]</url>
    <rssurl>[blog RSS URL]</rssurl>
    <atomurl>[blog Atom URL]</atomurl>
    <inboundblogs>[inbound blogs]</inboundblogs>
    <inboundlinks>[inbound links]</inboundlinks>
    <lastupdate>[date blog last updated]</lastupdate>
    <rank>[blog ranking]</rank>
    <lang></lang>
    <foafurl>[blog foaf URL]</foafurl>
  </weblog>
</result>
```

Digg API

- http://developers.digg.com/documentation/
- List Stories

API Query: **http://services.digg.com/1.0/endpoint?method=story.getAll&domain=nytimes.com**

Sample response:

Digg API


```
- <stories count="15" timestamp="1297897011" total="5595">
  + <story comments="509" diggs="518" href="http://digg.com/news/politics/climate_of_hate"
    id="20110110041444:b751731f-b2c8-49a4-b4c4-fc80c81bfad3"
    link="http://www.nytimes.com/2011/01/10/opinion/10krugman.html" media="0"
    promote_date="1294664468" status="top" submit_date="1294632884"></story>
  + <story comments="90" diggs="200" href="http://digg.com/news/politics
    /sarah_palin_s_nomination_chances_a_reassessment" id="20110101180146:3e86ef35-0c0e-
    4e0a-b34f-538392235554" link="http://fivethirtyeight.blogs.nytimes.com/2010/12/31/sarah-
    palins-nomination-chances-a-reassessment/" media="0" promote_date="1293982207"
    status="top" submit_date="1293904906"></story>
  + <story comments="175" diggs="532" href="http://digg.com/news/politics
    /the_war_on_logic" id="20110117023931:33a24afa-e0ab-4b99-8437-503a35e846fa"
    link="http://www.nytimes.com/2011/01/17/opinion/17krugman.html" media="0"
    promote_date="1295268682" status="top" submit_date="1295231971"></story>
  + <story comments="178" diggs="366" href="http://digg.com/news/politics/eat_the_future"
    id="20110214035028:8cf98e4d-19fb-44c7-ba90-9e0ef2adf7c3"
    link="http://www.nytimes.com/2011/02/14/opinion/14krugman.html" media="0"
    promote_date="1297686036" status="top" submit_date="1297655428"></story>
```

Digg API


- <stories count="15" timestamp="1297897011" total="5595">
+ <story comments="509" diggs="518" href="http://digg.com/news/politics/climate_of_hate"
lists 15 popular stories from
http://www.nytimes.com
id="20110214035028:8cf98e4d-19fb-44c7-ba90-9e0ef2adf7c3"
link="http://www.nytimes.com/2011/02/14/opinion/14krugman.html" media="0"
promote_date="1297686036" status="top" submit_date="1297655428"></story>

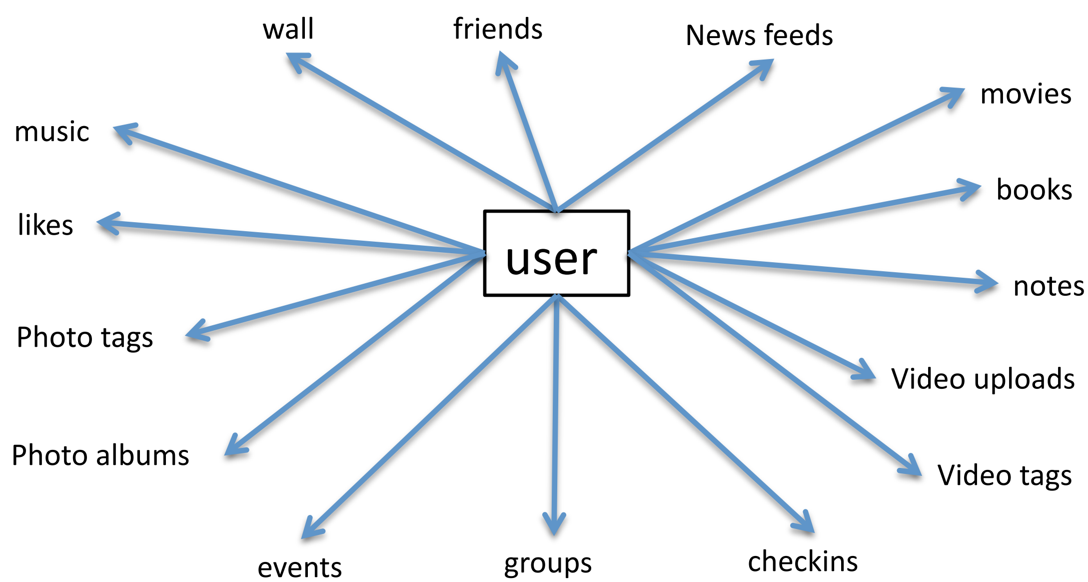
Facebook Graph API Structure

Objects

- Album
- Application
- Checkin
- Comment
- Event
- FriendList
- Group
- Insights
- Link
- Message
- Note
- Page
- Photo
- Post
- Status message
- Subscription
- Thread
- User
- Video

Each of these is an individual **object**

A user can also connect to these object. Facebook calls these as "**connections**"



Extensive documentation is available at:
(http://developers.facebook.com/docs/reference/api/)

Facebook Graph API Structure

- Objects
  - Album: A photo album
  - Application: An individual application registered on the Facebook Platform
  - Checkin: A check-in made through Facebook Places
  - Event: A Facebook event
  - Group: A Facebook group
  - Link: A shared link
  - Note: A Facebook note

Facebook Graph API Structure

- Objects
  - Page: A Facebook Page
  - Photo: An individual photo
  - Post: An individual entry in a profile's feed
  - Status message: A status message on a user's wall
  - Subscription: An individual subscription from an application to get real-time updates for an object type.
  - User: A user profile
  - Video: An individual video

User Object

http://developers.facebook.com/docs/reference/api/user/

Properties

`id`	The user's ID
`first_name`	The user's first name
`last_name`	The user's last name
`name`	The user's full name
`link`	A link to the user's profile
`about`	The user's blurb that appears under their profile picture
`birthday`	The user's birthday
`work`	A list of the work history from the user's profile
`education`	A list of the education history from the user's profile
`email`	The proxied or contact email address granted by the user
`website`	A link to the user's personal website.
`hometown`	The user's hometown
`location`	The user's current location
`bio`	The user's bio
`quotes`	The user's favorite quotes
`gender`	The user's gender
`interested_in`	Genders the user is interested in

Extended Permissions

http://developers.facebook.com/docs/authentication/permissions

User permission	Friends permission	Description
user_about_me	friends_about_me	Provides access to the "About Me" section of the profile in the about property
user_activities	friends_activities	Provides access to the user's list of activities as the activities connection
user_birthday	friends_birthday	Provides access to the full birthday with year as the birthday_date property
user_education_history	friends_education_history	Provides access to education history as the education property
user_events	friends_events	Provides access to the list of events the user is attending as the events connection
user_groups	friends_groups	Provides access to the list of groups the user is a member of as the groups connection
user_hometown	friends_hometown	Provides access to the user's hometown in the hometown property

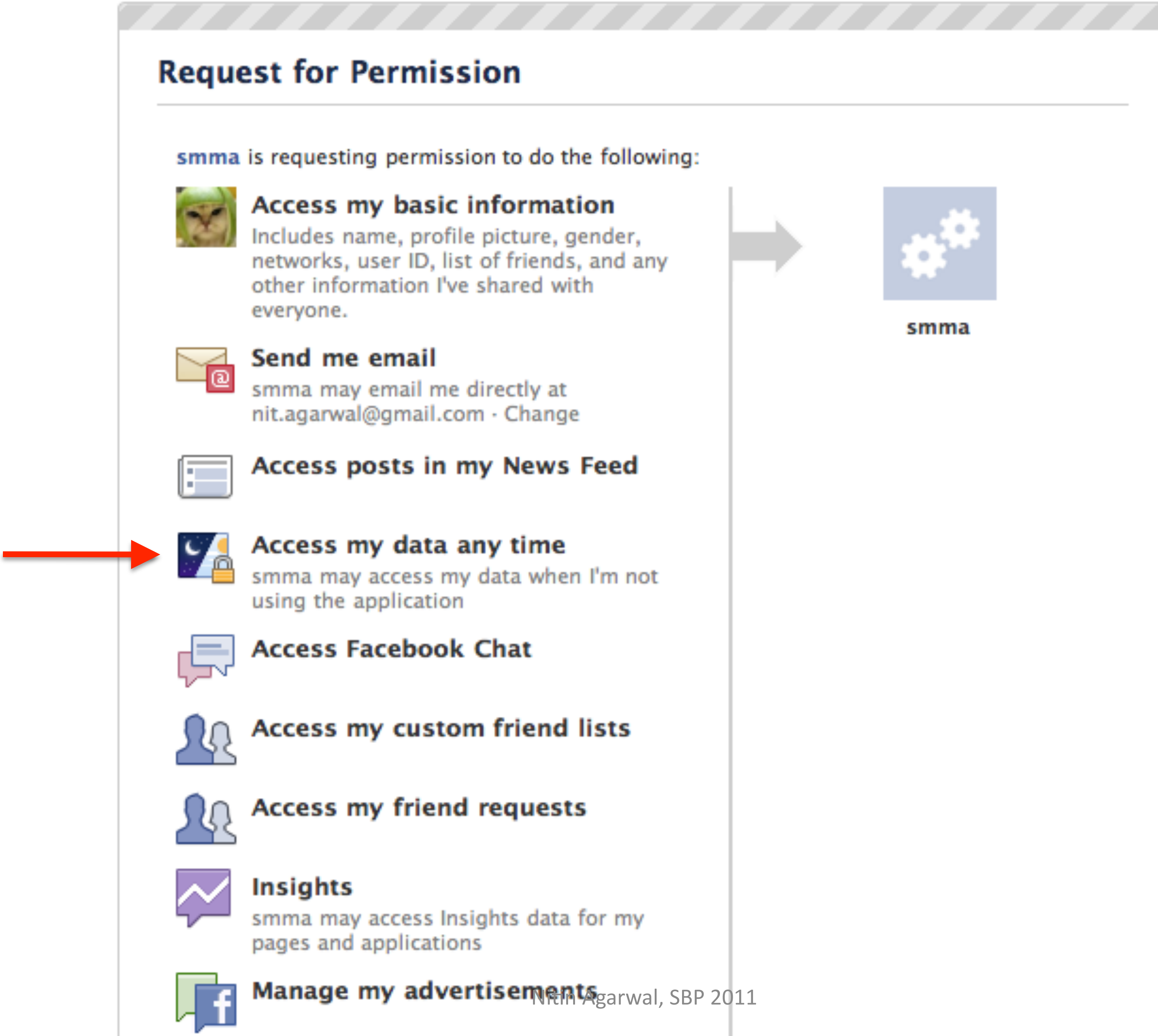
Data Permissions

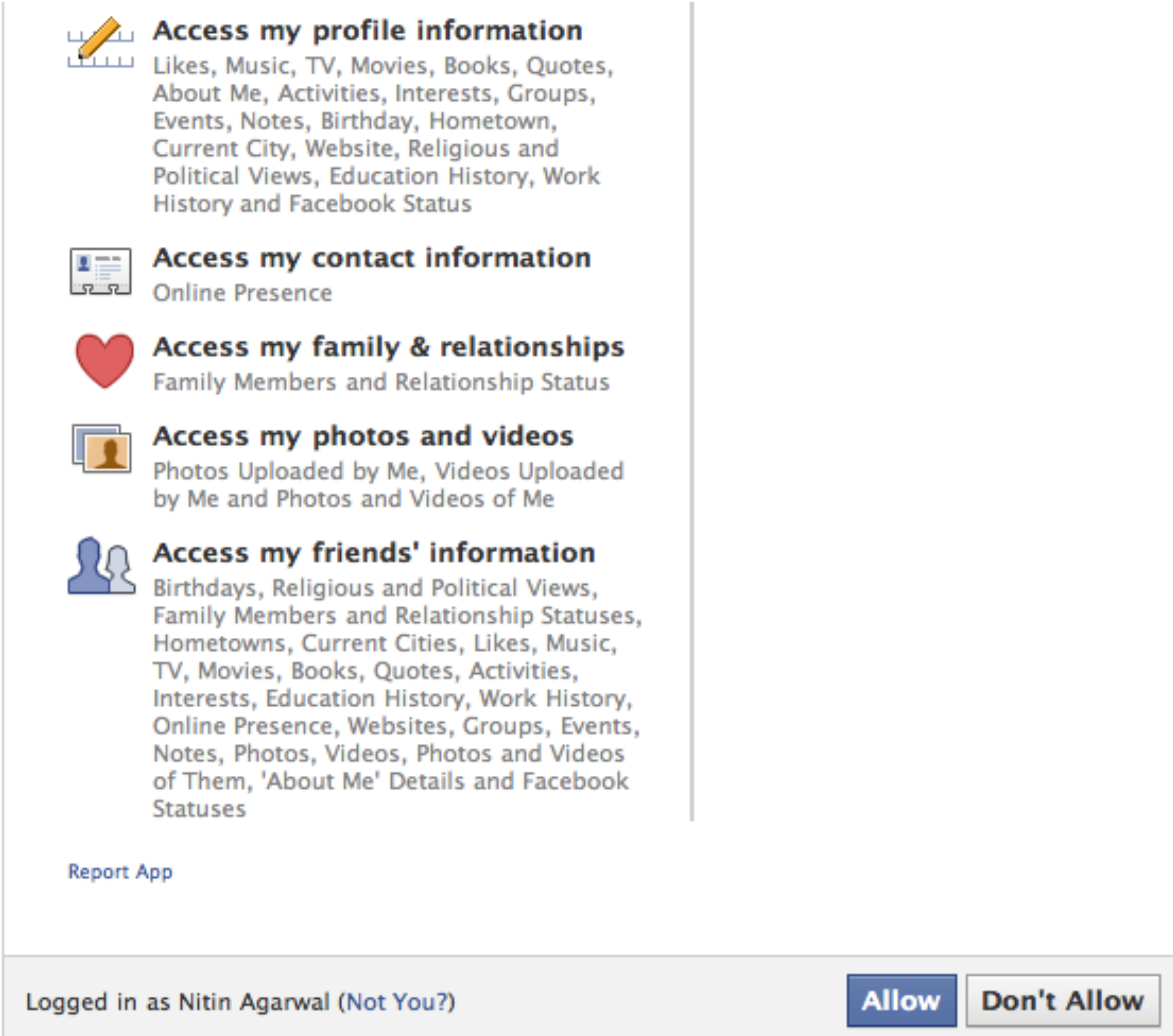
```
// login or logout url will be needed depending on current user state.
if ($me) {
  $logoutUrl = $facebook->getLogoutUrl();
} else {
$data_perms =
'offline_access,email,read_insights,read_stream,ads_management,xmpp_login,user_about_me,friends_about_me,user_activities,friends_activities,user_birthday,friends_birthday,user_education_history,friends_education_history,user_events,friends_events,user_groups,friends_groups,user_hometown,friends_hometown,user_interests,friends_interests,user_likes,friends_likes,user_location,friends_location,user_notes,friends_notes,user_online_presence,friends_online_presence,user_photo_video_tags,friends_photo_video_tags,user_photos,friends_photos,user_relationships,friends_relationships,user_religion_politics,friends_religion_politics,user_status,friends_status,user_videos,friends_videos,user_website,friends_website,user_work_history,friends_work_history,read_friendlists,read_requests';

  $loginUrl = $facebook->getLoginUrl(array('req_perms'=>$data_perms)); //with extended permission data objects
//  $loginUrl = $facebook->getLoginUrl(); //no extended permission data objects

//        'req_perms'       => 'email,read_insights',
//  $req_perms
//print_r($req_perms);
}
```

http://developers.facebook.com/docs/authentication/permissions


Request for Permission
smma is requesting permission to do the following:
Access my basic information
Includes name, profile picture, gender, networks, user ID, list of friends, and any other information I've shared with everyone.
Send me email
smma may email me directly at nit.agarwal@gmail.com · Change
Access posts in my News Feed
Access my data any time
smma may access my data when I'm not using the application
Access Facebook Chat
Access my custom friend lists
Access my friend requests
Insights
smma may access Insights data for my pages and applications
Manage my advertisements
smma


Access my profile information
Likes, Music, TV, Movies, Books, Quotes, About Me, Activities, Interests, Groups, Events, Notes, Birthday, Hometown, Current City, Website, Religious and Political Views, Education History, Work History and Facebook Status
Access my contact information
Online Presence
Access my family & relationships
Family Members and Relationship Status
Access my photos and videos
Photos Uploaded by Me, Videos Uploaded by Me and Photos and Videos of Me
Access my friends' information
Birthdays, Religious and Political Views, Family Members and Relationship Statuses, Hometowns, Current Cities, Likes, Music, TV, Movies, Books, Quotes, Activities, Interests, Education History, Work History, Online Presence, Websites, Groups, Events, Notes, Photos, Videos, Photos and Videos of Them, 'About Me' Details and Facebook Statuses
Report App
Logged in as Nitin Agarwal (Not You?)
Allow
Don't Allow

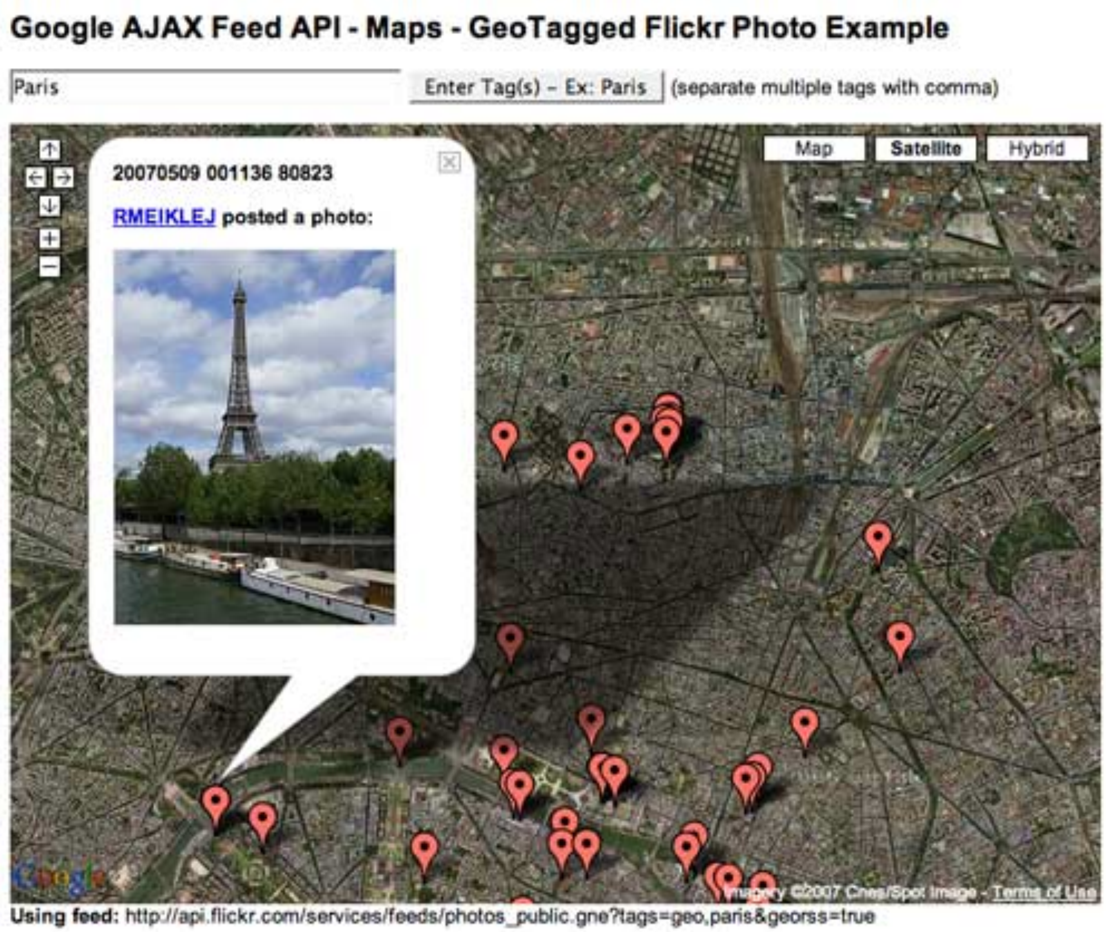
Other Similar APIs

- Google Social Graph API
  - http://code.google.com/apis/socialgraph/
  - Extensive documentation at: http://code.google.com/apis/socialgraph/docs/
- MySpace OpenSocial API
  - http://wiki.developer.myspace.com/index.php?title=OpenSocial_Applications
  - Short 20 minute video: http://myspacetv.com/index.cfm?fuseaction=vids.individual&videoid=26838901
  - Tutorials: http://wiki.developer.myspace.com/index.php?title=OpenSocial_Version_1.0

Mashup – Example - 1

- Find most popular stories of a blog site. For each of these blog posts get the top 5 tags.

Flickr-Google Mashup - 2



Some Available Datasets

- Nielsen Buzzmetrics dataset (http://www.icwsm.org/format.txt)
  - ~ 14M blog posts from 3M blog sites collected by Nielsen BuzzMetrics in May 2006
  - 1.7M blog-blog links
  - Up to a half of the blog outlinks are missing
  - 51% of the total blog posts are in English
- Enron Email dataset (http://www.cs.cmu.edu/~enron/)
  - Emails from about 150 users
  - The corpus contains a total of about 0.5M messages
  - People have studied the social networks between users based on link construction
  - Links are constructed based on email senders and recipients

Available Datasets (2)

- TREC (http://ir.dcs.gla.ac.uk/test_collections/blog06info.html)
  - A crawl of Feeds, and associated Permalink and homepage documents (from late 2005 and early 2006)
  - 100,649 feeds were polled once a week for 11 weeks
  - Total Number of Feeds collected:753,681
  - Average feeds collected every day:10,615
  - Uncompressed Size:38.6GB Compressed Size:8.0GB
  - Reasonably sized spam component for added realism
  - Fee: £400 ~ $794.36

Social Media

- Organic, opinionated, open-source (quite a bit) data
- Social networks
  - Involving people, places, products, organizations
  - Communities
  - Interactions
- Computing
  - Modeling
  - Mining
  - Prediction

DATA ANALYSIS

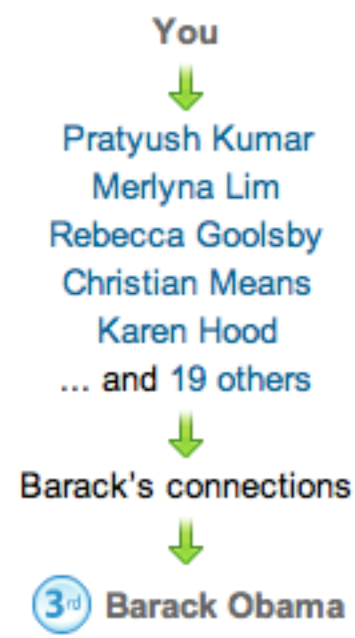
Measures, Models, and Methods

- Graph theoretic analysis
  - SNA/Centrality Measures
  - Random, scale-free, preferential attachment, hybrid, cascade
  - Link analysis
- Content analysis techniques
  - Supervised/unsupervised learning algorithms

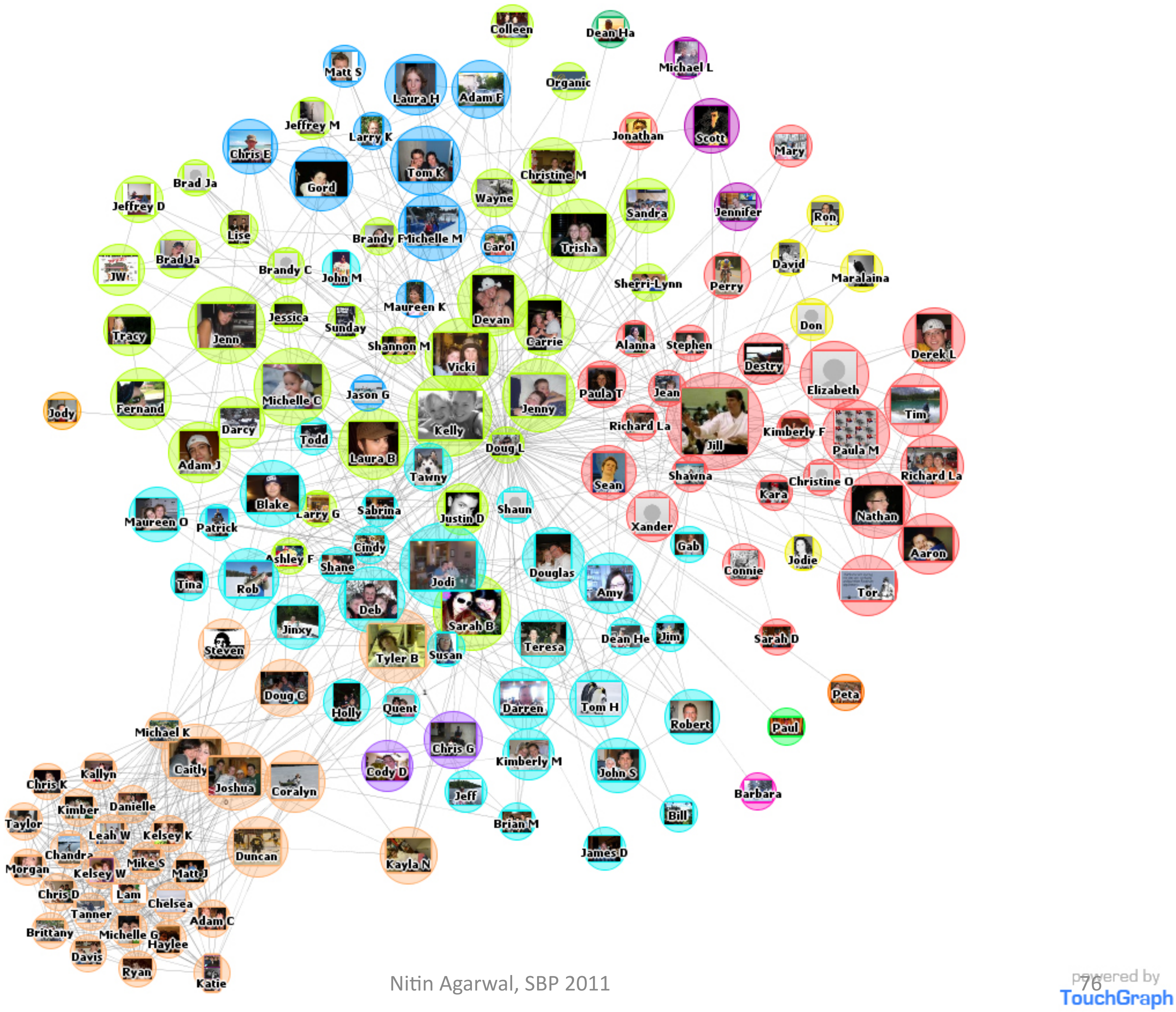
Applications of Graph Theory

- Social Networks
  - Example: Model relationships between social entities like individuals, groups
  - Sample Analysis: Who are the most centrally connected people? Is there a path between two users? What is the average path-length?

    Example: I know ‘someone’ who knows ‘someone’ who knows Barack Obama (LinkedIn)!



- Blogosphere
  - Vertices : Bloggers/Blog posts/Blog sites
  - Edges: Relationships/Links

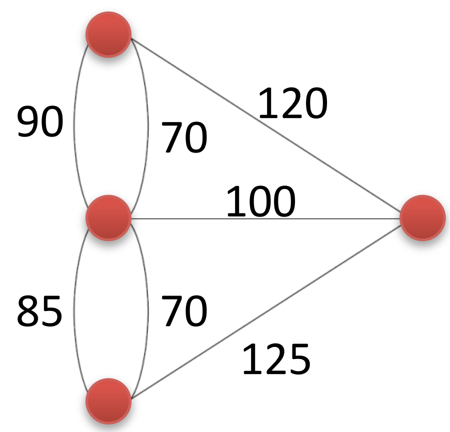

Colleen
Dean Ha
Michael L
Matt S
Laura H
Adam F
Organic
Jeffrey M
Larry K
Jonathan
Scott
Mary
Chris E
Tom K
Christine M
Gord
Brad Ja
Jeffrey D
Wayne
Sandra
Jennifer
Ron
Lise
Brandy F
Michelle M
Carol
Trisha
Brad Ja
JW
Brandy C
John M
David
Maralaina
Sherri-Lynn
Perry
Maureen K
Devan
Jessica
Sunday
Carrie
Don
Tracy
Jenn
Shannon M
Alanna
Stephen
Destry
Derek L
Vicki
Elizabeth
Jody
Fernand
Michelle C
Jason G
Paula T
Jean
Jenny
Tim
Darcy
Kelly
Richard La
Kimberly F
Paula M
Todd
Doug L
Jill
Adam J
Laura B
Tawny
Shawna
Richard La
Christine O
Sean
Kara
Blake
Larry G
Sabrina
Justin D
Shaun
Nathan
Maureen O
Patrick
Xander
Cindy
Gab
Jodie
Aaron
Ashley F
Shane
Douglas
Connie
Tina
Rob
Jodi
Amy
Tor
Deb
Jinxy
Sarah B
Steven
Dean He
Jim
Sarah D
Tyler B
Susan
Teresa
Peta
Doug C
Holly
Quent
Darren
Tom H
Robert
Paul
Michael K
Chris G
Kimberly M
Caitly
Chris K
Kallyn
Joshua
Coralyn
Cody D
John S
Barbara
Jeff
Kimber
Danielle
Taylor
Bill
Brian M
Leah W
Kelsey K
Duncan
Chandra
James D
Mike S
Kayla N
Morgan
Kelsey W
Matt J
Chris D
Lam
Chelsea
Tanner
Adam C
Brittany
Michelle G
Haylee
Davis
Ryan
Katie
powered by
TouchGraph

Types of Graphs

- **Directed/Undirected Graphs**
  - Each edge has an orientation/direction
  - To model asymmetrical relationships (like one-way street, etc.)
  - Undirected graph is a special case of Directed graph, in which every edge can be split into two directed edge in either directions
  - Directed graph referred to as "Digraph", most often "Graph" simply means undirected
  - Examples?

Types of Graphs

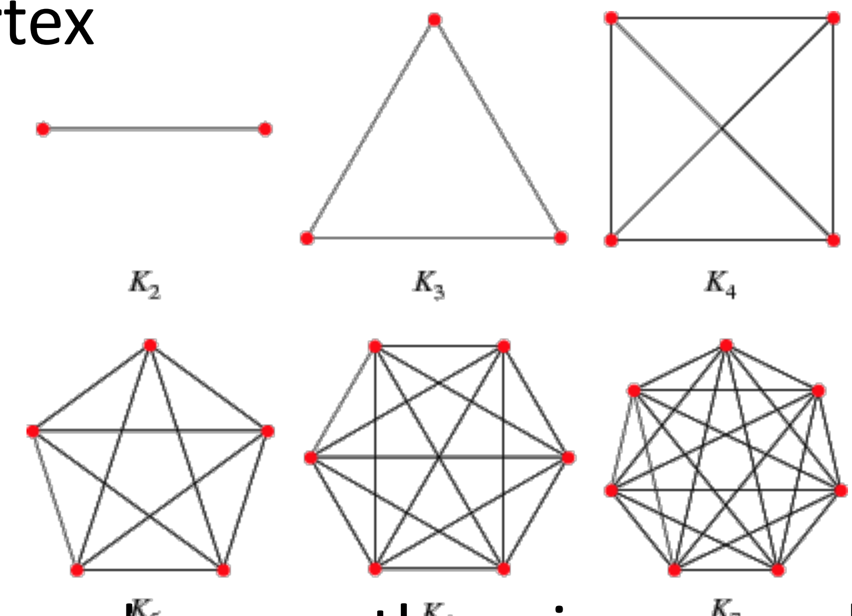
- **Weighted Graph**
  - Weight w(u,v) on each edge
  - Can model things like distance, delay, bandwidth, importance of an edge, etc.



  - Unweighted graph is a special case of weighted graph with w(u,v) = 1 for all edges

Types of Graphs

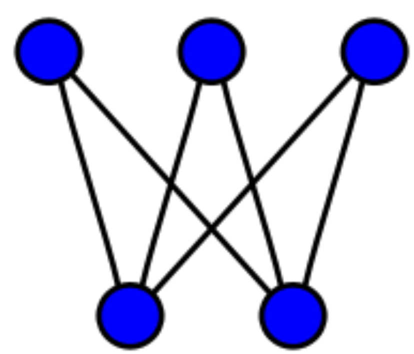
- **Complete Graph** (aka Clique)
  - A graph in which each vertex is connected to every other vertex



  - How many edges are there in a complete graph of *n* nodes. (n*n-1)/2

Types of Graphs

- Bipartite graphs
  - Two clusters of nodes, such that there is no edge between nodes of the same cluster.
  - There are edges between nodes belonging to different clusters.



Types of Graphs

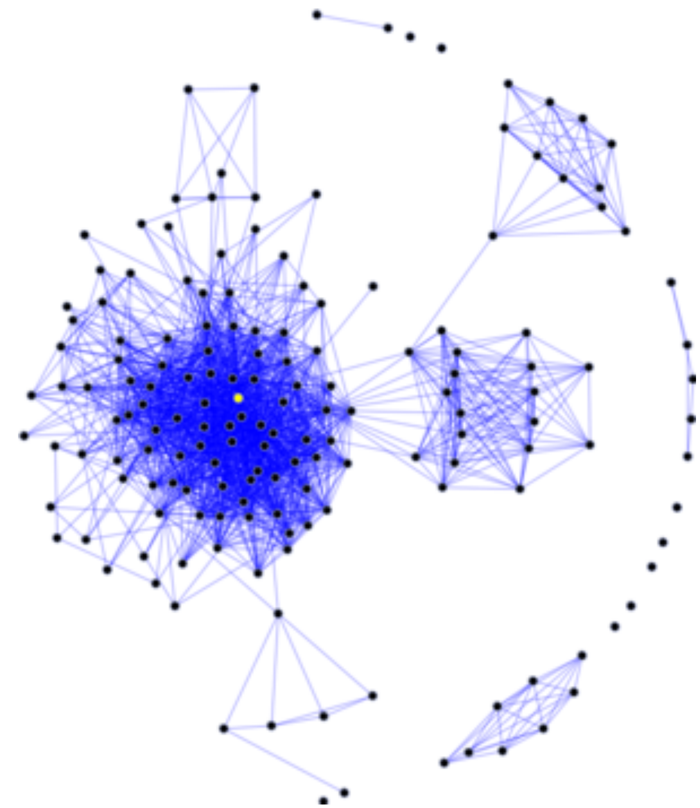
- Homogeneous graph
  - All the vertices are of same type
  - Also, 1-mode graph, social network of users
- Heterogeneous graph
  - Vertices could be of different types
  - 2-mode (two different types of nodes), examples?
  - 3-mode
  - …

Types of Graphs

- **Subgraph** of a graph G=(V,E)
  - Graph G'=(V',E')
    - whose vertex set V' is subset of V
    - edge set is a subset of E restricted to vertices in V'
- **Supergraph** of a graph G'
  - Graph in which G' is a subgraph
- These concepts are extremely

  helpful in defining communities

  of users in social networks
- **Spanning Subgraph** (**Factor**) of a graph G=(V,E)
  - Subgraph that contains (spans) all vertices of G



Graphs: Definitions

- **Connected Component**
  - A connected component or simply component of a graph is a maximal subgraph in which all vertices of the subgraph are reachable from each other
- **Strongly Connected Component** of Digraph
  - Maximal subgraph in which all vertices of the subgraph are reachable from each other following the directions of the edges

Graphs: Definitions

- **Distance** between two vertices: $d_G(u,v)$
  - Length of the shortest path between u and v in G
  - Special cases:

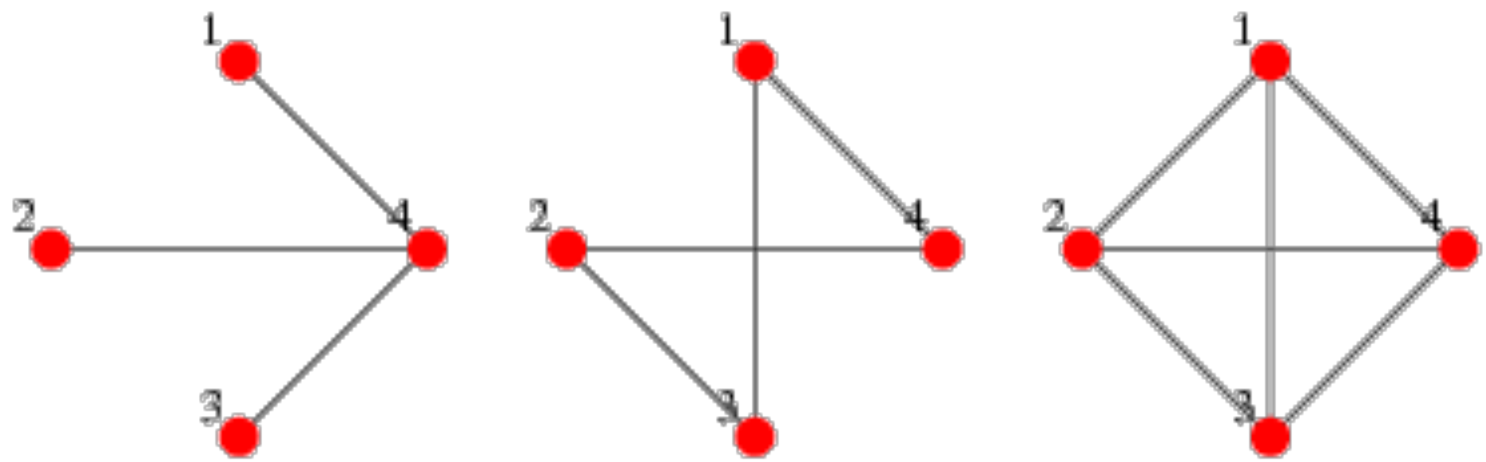
    When $u = v$, $d_G(u,v) = 0$

    When u and v are unreachable, $d_G(u,v) = \infty$
- **Diameter** of a graph
  - Largest *Distance* between any pair of vertices in the graph

Sociometry

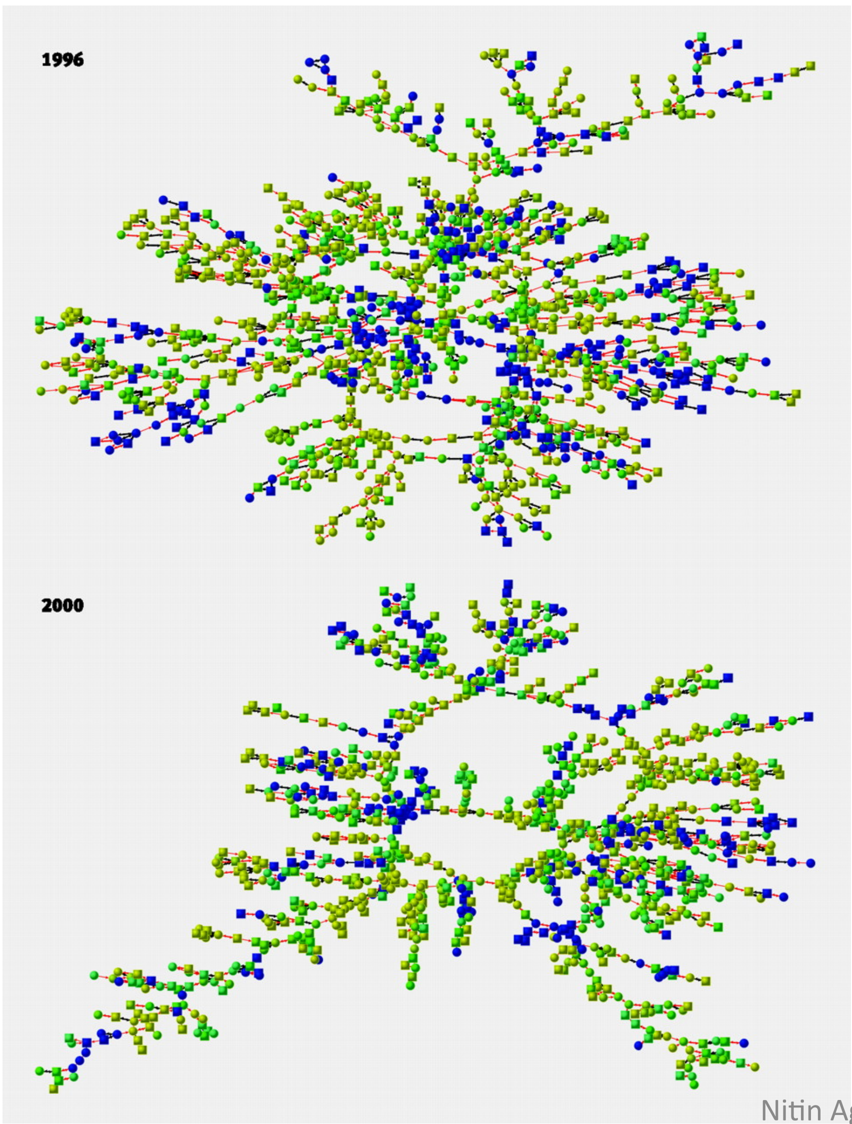
- Sociogram
  - Points in a sociogram that can make choices are called *actors* (compared to nodes in graphs). Isolates are those that have few or no choices.
  - Actors that chose each other (make mutual choices) connect with edge → undirected graph
  - One-way choices → directed graph, choices are not reciprocated (example?)
  - Groups of actors that chose each other → Cliques
- Ego-alter network
  - In your network of friends, you are the "ego" and your friends are the "alters"

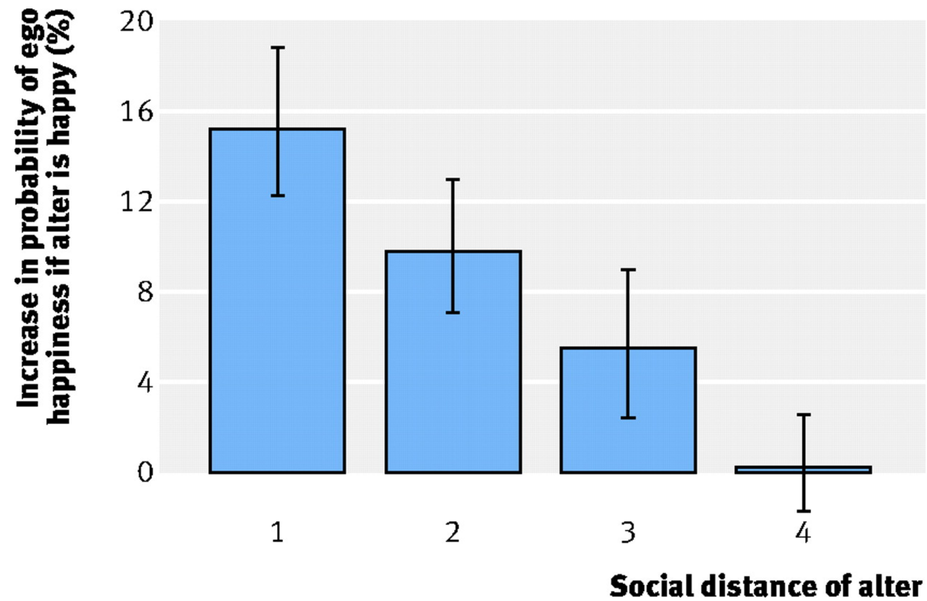
Sociogram

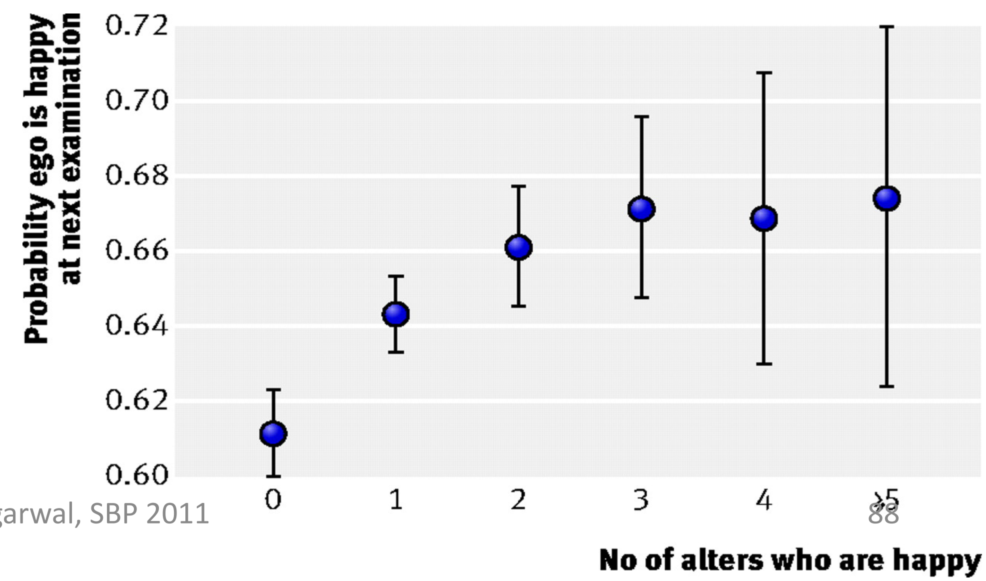


$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

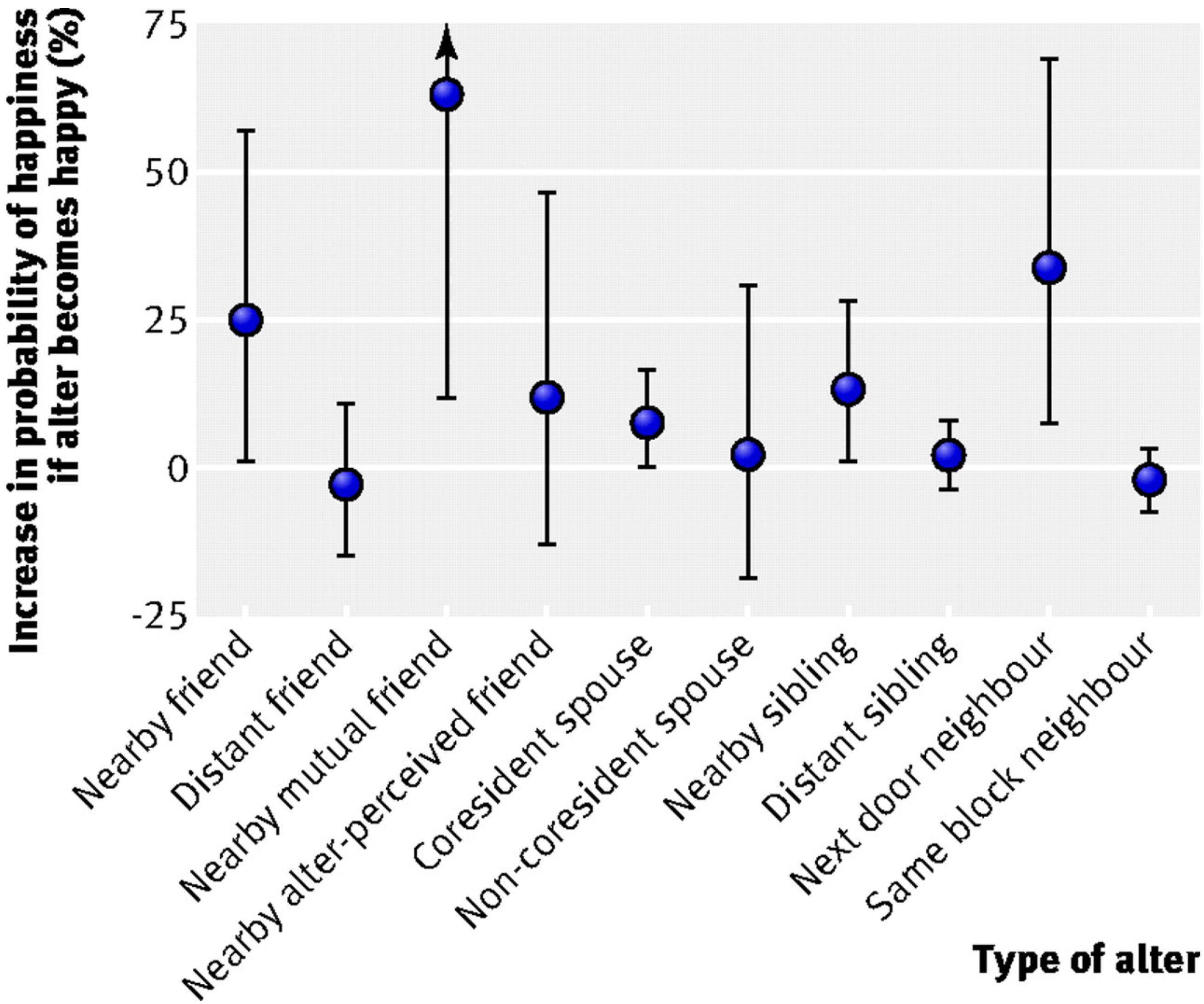
Spread of Happiness







Spread of Happiness



Source: http://www.bmj.com/content/337/bmj.a2338.full

Other Virtues

- Spread of Smoking
  - (http://jhfowler.ucsd.edu/collective_dynamics_of_smoking.pdf)
- Spread of Obesity
  - (http://jhfowler.ucsd.edu/spread_of_obesity.pdf)

Centrality Measures

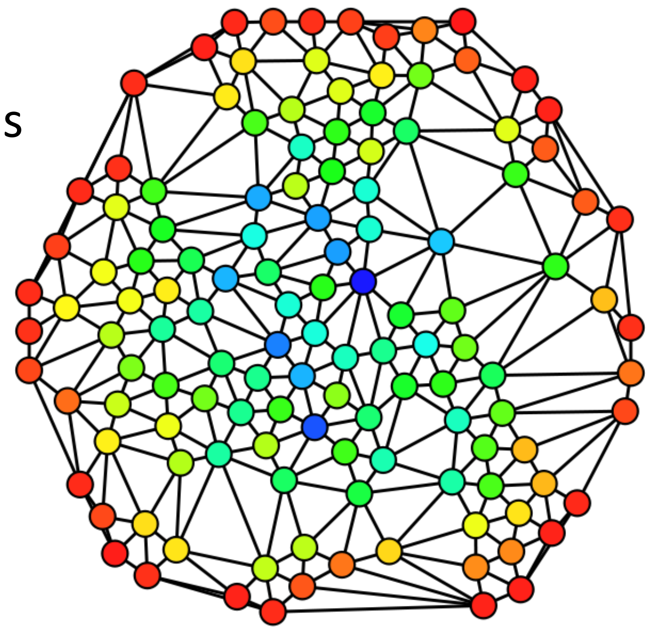
- Degree centrality
  - Defined as the number of ties a node has

$$C_d(v) = \left|\{e : M_{adj}(v, v_j) \neq 0, \forall j\}\right|$$

  - For directed network
    - Indegree ~ "popularity"
    - Outdegree ~ "gregariousness"
  - $O(V^2)$ for V vertices in dense network
  - O(E) for E edges in sparse network

Centrality Measures

- Betweenness centrality
  - a centrality measure of a vertex within a graph
  - Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not
  - Act as “broker” or “bridge”
  - $O(V^3)$ complexity
  - $O(V^2 \log V + VE)$ for sparse network



$$C_B(v) = \sum_{\substack{s \neq v \neq t \in V \\ s \neq t}} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

σ_{st} is the geodesic path between s and t. $\sigma_{st}(v)$ is the geodesic path between s and t passing through v.

Centrality Measures

- Closeness centrality
  - A centrality measure of a vertex within a graph
  - Vertices that tend to have short geodesic distances to other vertices within the graph have higher closeness.
  - Defined as the mean geodesic distance between a vertex v and all other reachable vertices

$$\frac{\sum_{t \in V \setminus v} d_G(v,t)}{n-1}$$

  - $O(V^3)$ complexity

Centrality Measures

- Eigenvector centrality
  - Measure of the importance of a node in a network
  - Assigns relative scores to all nodes in the network
  - Better to connect to more “popular” nodes than less “popular” ones
  - Google's PageRank is a variant of the Eigenvector centrality measure

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{N} A_{i,j} x_j \quad \text{or} \quad \vec{x} = \frac{1}{\lambda} A\vec{x}$$

Prestige

- Prestige is a more refined measure of prominence of an actor than centrality.
  - Difference: ties sent (out-links) and ties received (in-links).
- A prestigious actor is one who is a recipient of several ties.
  - To compute the prestige: we use only in-links.
- Difference between centrality and prestige:
  - centrality focuses on all the links
  - prestige focuses only on in-links.

Different Prestige Measures

- We study three prestige measures.
  - Degree Prestige
  - Proximity Prestige
  - Rank Prestige
- **Rank prestige** forms the basis of most Web page link analysis algorithms, including PageRank and HITS.

Degree prestige

Based on the definition of the prestige, it is clear that an actor is prestigious if it receives many in-links or nominations. Thus, the simplest measure of prestige of an actor i (denoted by $P_D(i)$) is its in-degree.

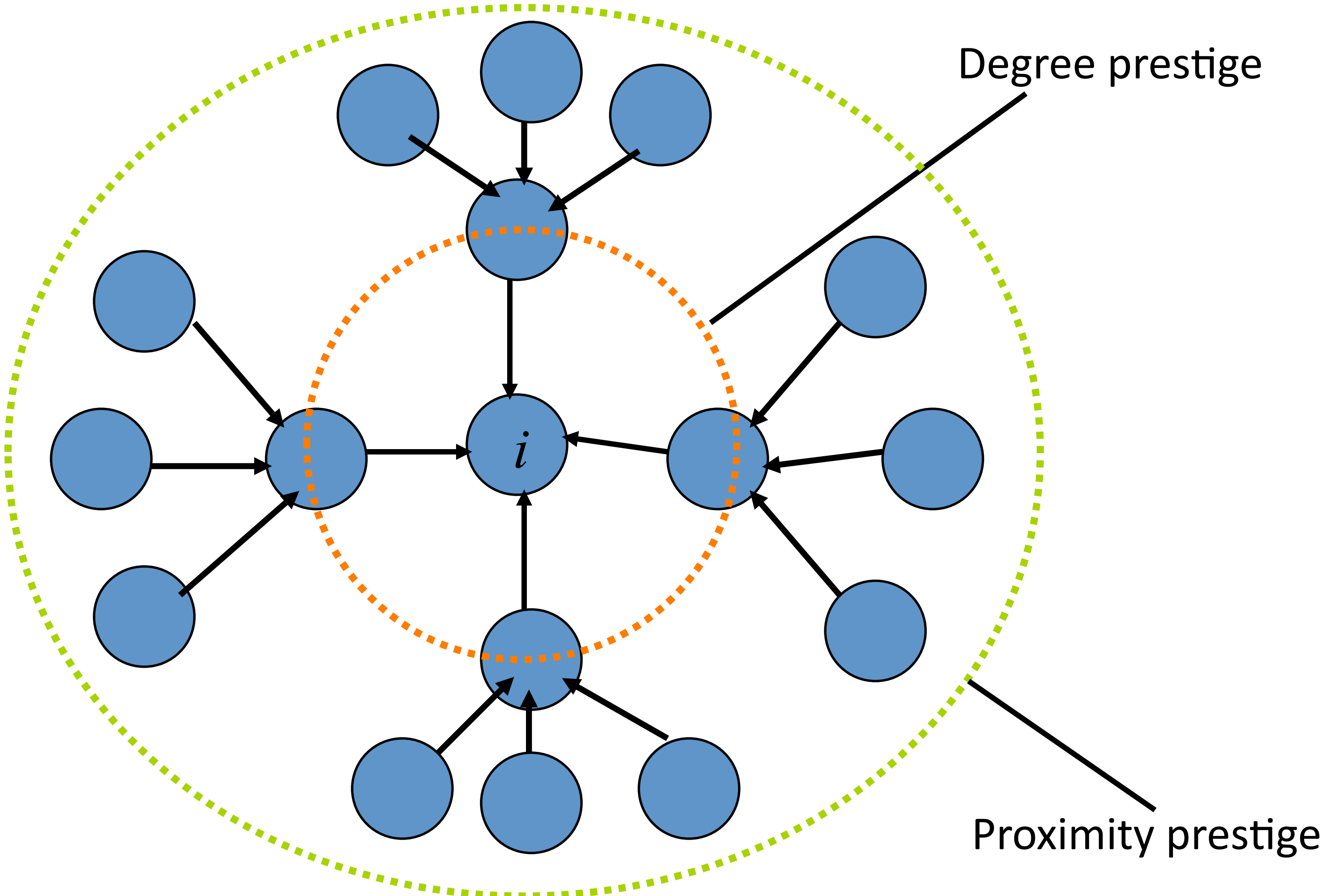
$$P_D(i) = \frac{d_I(i)}{n-1}, \quad (6)$$

where $d_I(i)$ is in-degree of i (the number of in-links of actor i_i) and n is the total number of actors in the network. As in the degree centrality, dividing $n - 1$ standardizes the prestige value to the range from 0 and 1. The maximum prestige value is 1 when every other actor links to or chooses actor i.

Proximity prestige

- The degree prestige of an actor *i* only considers the actors that are adjacent to *i*.
- The proximity prestige generalizes it by considering both the actors directly and indirectly linked to actor *i*.
  - We consider every actor *j* that can reach *i*.
- Let I_i be the set of actors that can reach actor *i*.
- The **proximity** is defined as closeness in terms of distance of other actors to *i*.
- Let $d(j, i)$ denote the distance from actor *j* to actor *i*.

Degree vs. Proximity Prestige



Domain of Influence

Rank prestige

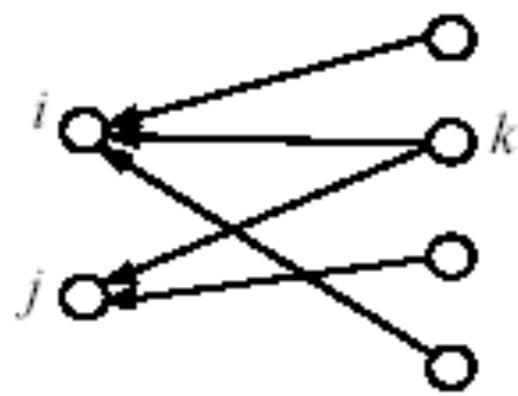
- In the previous two prestige measures, an important factor is not considered,
  - the **prominence** of individual actors who do the "voting"
- E.g., A webpage that is linked by New York Times is more prestigious than if it is linked by some arbitrary website
- If one's circle of influence is full of prestigious actors, then one's own prestige is also high.
  - Thus one's prestige is affected by the ranks or statuses of the involved actors.

Co-citation and Bibliographic Coupling

- Another area of research concerned with links is **citation analysis** of scholarly publications.
  - A scholarly publication cites related work to acknowledge the origins of some ideas and to compare the new proposal with existing work.
- When a paper cites another paper, some relationship can be derived between the publications.
  - *If two papers **are cited** by the same papers, they are related*
  - *If two papers **cite** the same papers, they are also related*
- We discuss two types of citation analysis, **co-citation** and **bibliographic coupling**.

Co-citation

- If articles *i* and *j* are both cited by article *k*, then they may be related in some sense to one another.
- The more articles they are cited by, the stronger their similarity is.



Co-citation

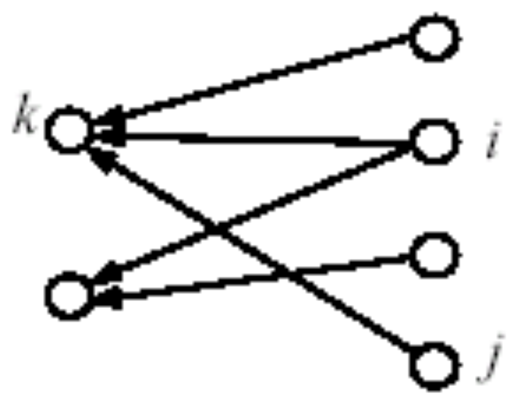
- Let ***L*** be the citation matrix. Each cell of the matrix is defined as follows:
  - L_{ij} = 1 if article *i* cites article *j*, and 0 otherwise.
- **Co-citation** (denoted by C_{ij}) is a similarity measure defined as the number of articles that co-cite *i* and *j*,

$$C_{ij} = \sum_{k=1}^{n} L_{ki} L_{kj}$$

- ***C*** is symmetric

Bibliographic coupling

- Bibliographic coupling operates on a similar principle.
- Bibliographic coupling considers articles that cite the same articles
  - if articles *i* and *j* both cite article *k*, they may be related/similar.
- The more articles they both cite, the stronger their similarity is.

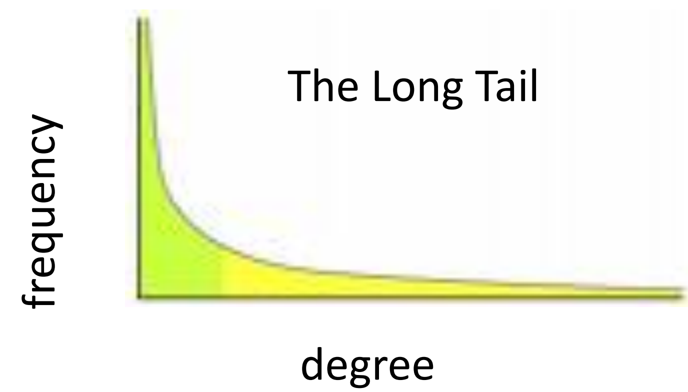


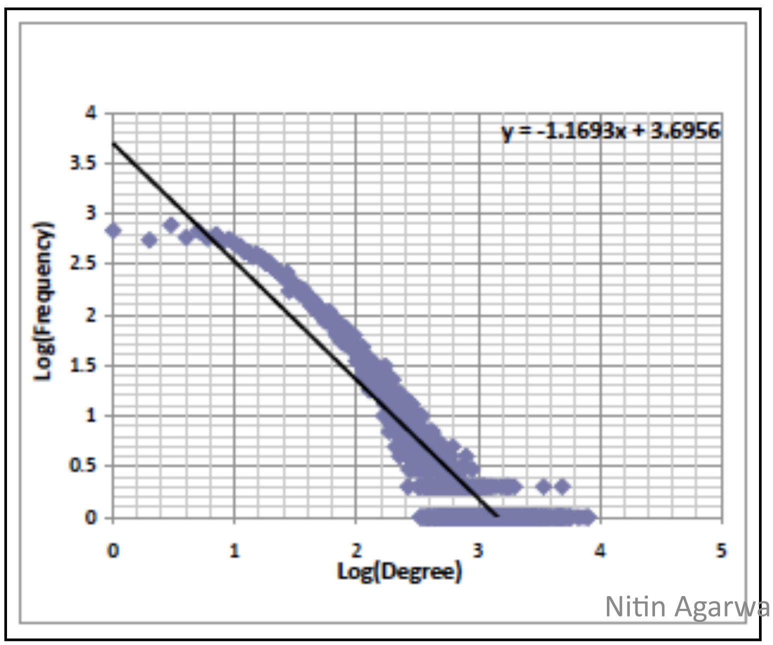
$$B_{ij} = \sum_{k=1}^{n} L_{ik} L_{jk}$$

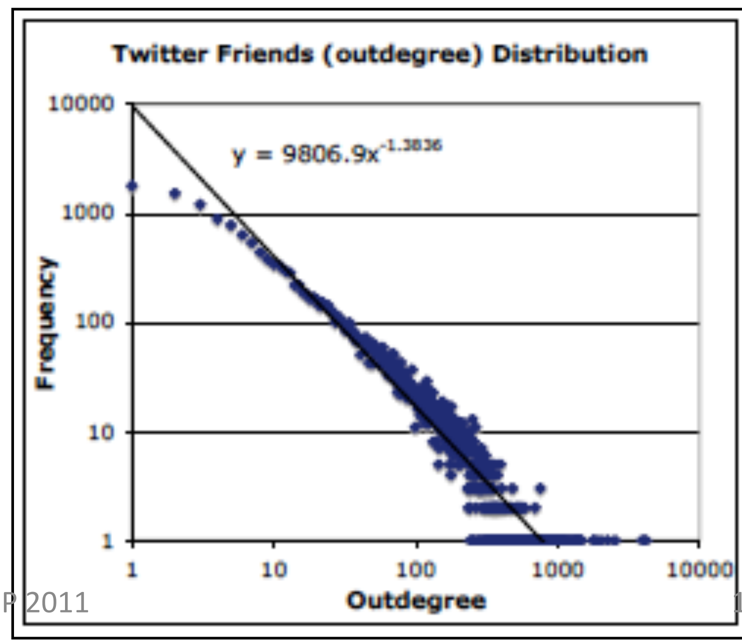
B is symmetric

Social Networks

- Power law degree distribution
- $f(x) = ax^{-ß}$
- $\log(f(x)) = \log(a) - ß \log(x)$

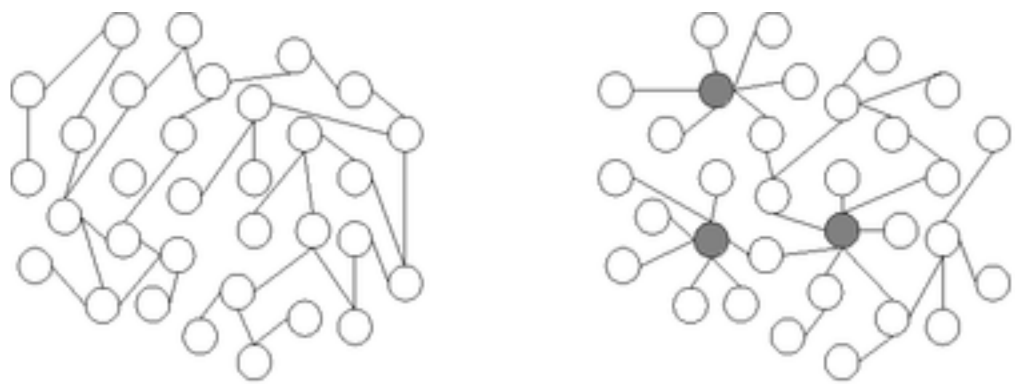






Social Networks

- Scale-free networks
  - 2 < ß < 3



(a) Random network (b) Scale-free network

- Preferential attachment model
  - "Rich get richer" effect

$$P(e_{ij}) \propto \frac{d(v_i)}{|V|}$$

Undirected graph

$$P(e_{i \leftarrow j}) \propto \frac{d_{in}(v_i)}{|V|} \quad P(e_{i \rightarrow j}) \propto \frac{d_{out}(v_i)}{|V|}$$

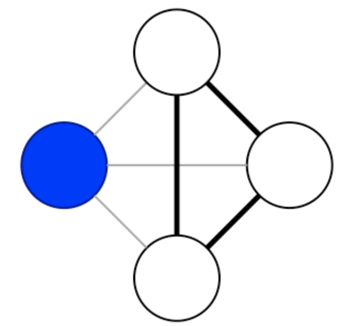
Directed graph

Social Networks

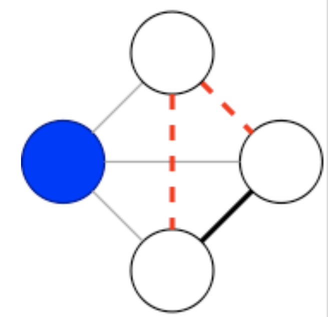
- Clustering coefficient

$$C = \frac{\text{number of closed triplets}}{\text{number of triplets}}.$$

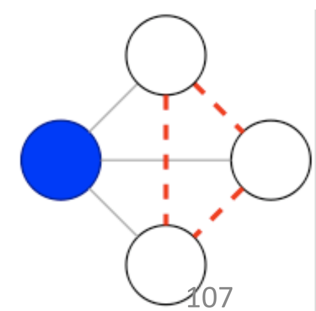
- Cliquishness
- **Social networks have larger clustering coefficient values as compared to random network**



c = 1



c = 1/3



Graph based Clustering

- Spectral clustering

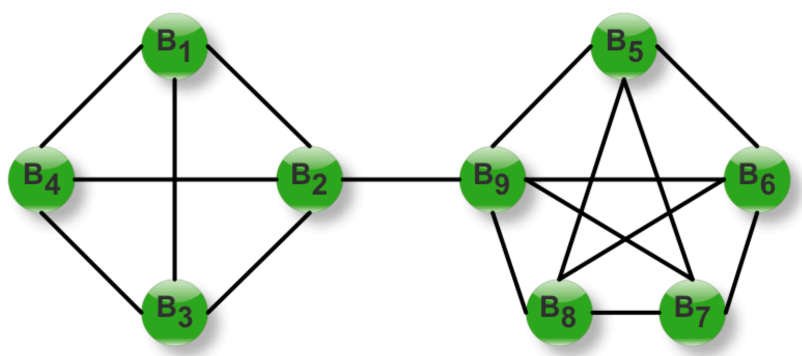
Input : Adjacency matrix: W,
Number of clusters: k
Output: k clusters of n nodes/blogs in the blog graph

1 Compute the diagonal matrix, D;
2 Compute the graph laplacian, $L = D - W$;
3 Compute the first k eigenvectors, $e_1, e_2, ..., e_k$ of L;
4 Juxtapose these eigenvectors to construct a $n \times k$ matrix;
5 Compute k clusters using k-means algorithm on this matrix;

Algorithm 1: Algorithm for spectral clustering.

Graph based Clustering

- Spectral clustering example



(a) Blog Network

	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9
B_1	0	1	1	1	0	0	0	0	0
B_2	1	0	1	1	0	0	0	0	1
B_3	1	1	0	1	0	0	0	0	0
B_4	1	1	1	0	0	0	0	0	0
B_5	0	0	0	0	0	1	1	1	1
B_6	0	0	0	0	1	0	1	1	1
B_7	0	0	0	0	1	1	0	1	1
B_8	0	0	0	0	1	1	1	0	1
B_9	0	1	0	0	1	1	1	1	0

(b) Matrix: W

Graph based Clustering

- Spectral clustering example

	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9
B_1	3	0	0	0	0	0	0	0	0
B_2	0	4	0	0	0	0	0	0	0
B_3	0	0	3	0	0	0	0	0	0
B_4	0	0	0	3	0	0	0	0	0
B_5	0	0	0	0	4	0	0	0	0
B_6	0	0	0	0	0	4	0	0	0
B_7	0	0	0	0	0	0	4	0	0
B_8	0	0	0	0	0	0	0	4	0
B_9	0	0	0	0	0	0	0	0	5

(c) Matrix: D

	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9
B_1	3	-1	-1	-1	0	0	0	0	0
B_2	-1	4	-1	-1	0	0	0	0	-1
B_3	-1	-1	3	-1	0	0	0	0	0
B_4	-1	-1	-1	3	0	0	0	0	0
B_5	0	0	0	0	4	-1	-1	-1	-1
B_6	0	0	0	0	-1	4	-1	-1	-1
B_7	0	0	0	0	-1	-1	4	-1	-1
B_8	0	0	0	0	-1	-1	-1	4	-1
B_9	0	0	0	0	-1	-1	-1	-1	5

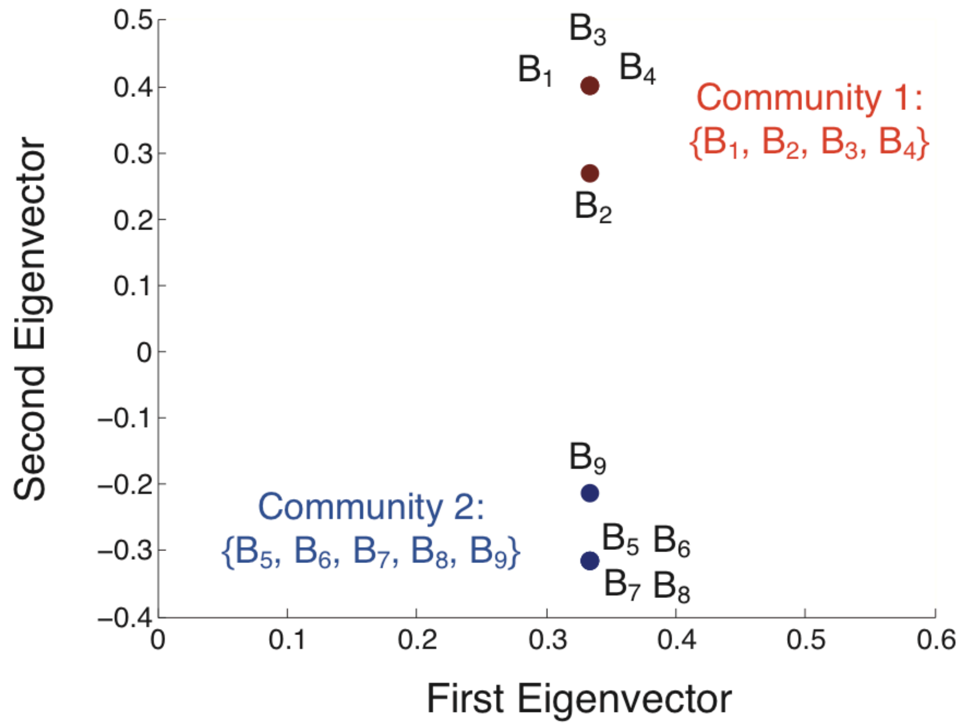
(d) Matrix: L $(=D\text{-}W)$

Graph based Clustering

- Spectral clustering example

	EV_1	EV_2
B_1	0.3333	0.4015
B_2	0.3333	0.2701
B_3	0.3333	0.4015
B_4	0.3333	0.4015
B_5	0.3333	-0.3156
B_6	0.3333	-0.3156
B_7	0.3333	-0.3156
B_8	0.3333	-0.3156
B_9	0.3333	-0.2123

(e) First two eigenvectors of L



(f) Visualization

Content Analysis Techniques

- Social media have rich textual content
- Not only people create new content, they also enrich the existing content by providing meta data such as labels and tags
- Human-generated tags are also called folksonomies
- State-of-the-art content analysis techniques could be used for basic clustering, classification of the blog posts/blog sites

Content Analysis Techniques

- *tf-idf* could be used for indexing the text
- Folksonomies could be considered as class labels
- Supervised machine learning
  - Predict tags of unlabeled corpus
  - Predict links
- Spam classification
- Topic modeling (LDA) could be used to identify off-topic chatter

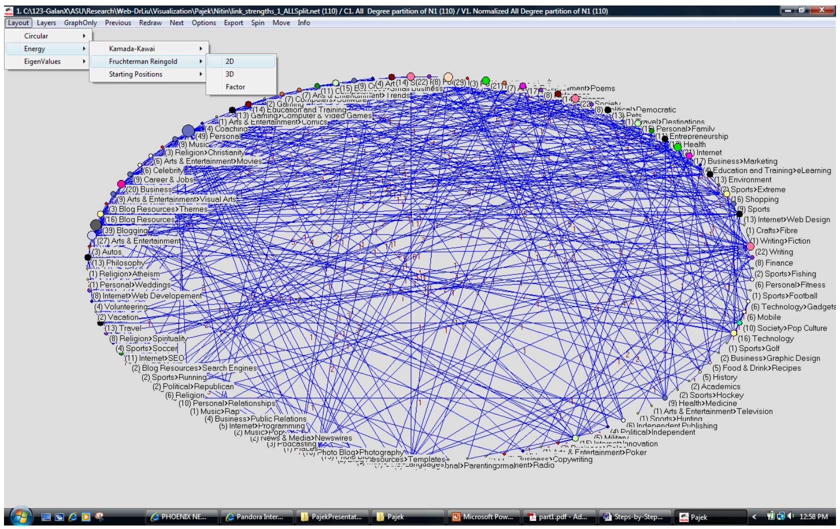
Visual Analytics

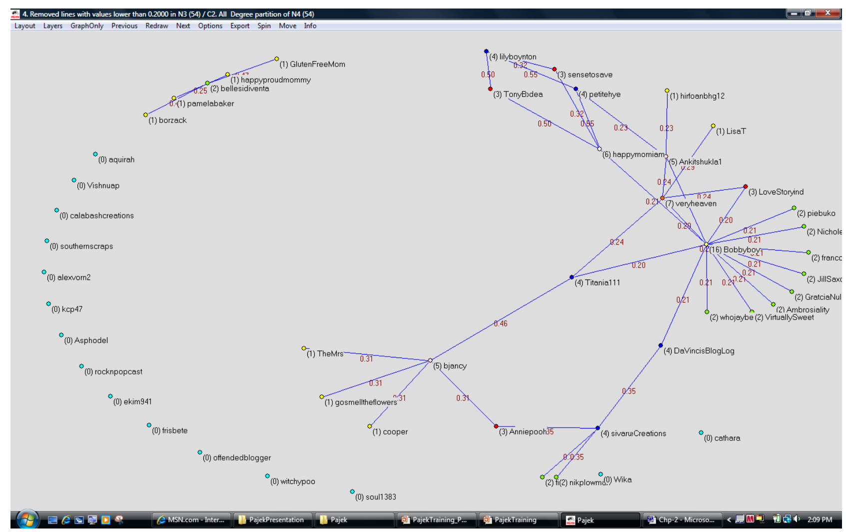
- Technique to graphically represent sets of data.
- Helps make the data easier to read or understand.
- Preprocess the data
- Zoom-in directly to the points-of-interests.
- In addition to good visualization
  - Statistical and analytical capabilities
  - Data structuring, clustering, labeling, classification…

Pajek Description

- Pajek is a Windows based program for analyzing large networks.
  - Developed by Vladimir Batagelj and Andrej Mrvar from University of Ljubljana
- It is **FREE**!
  - Freeware software can be downloaded from
    - http://vlado.fmf.uni-lj.si/pub/networks/pajek/
- It has a comprehensive manual focused on Social Network Analysis techniques & metrics.
  - Integrates Theory, Applications and the Software.
  - Link to book, images and samples:
    - http://vlado.fmf.uni-lj.si/pub/networks/book/

Simplify the Networks



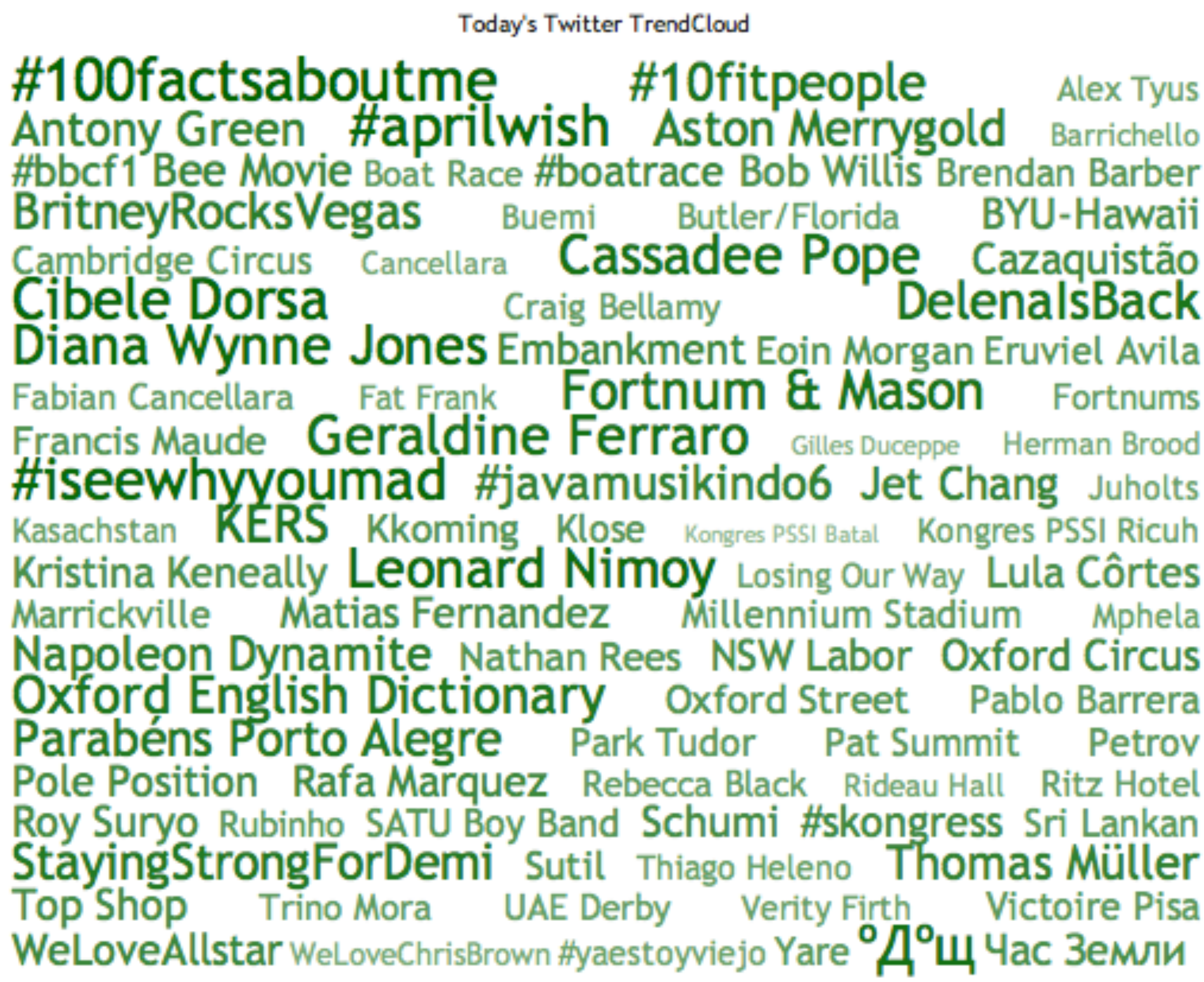


Content Visualization

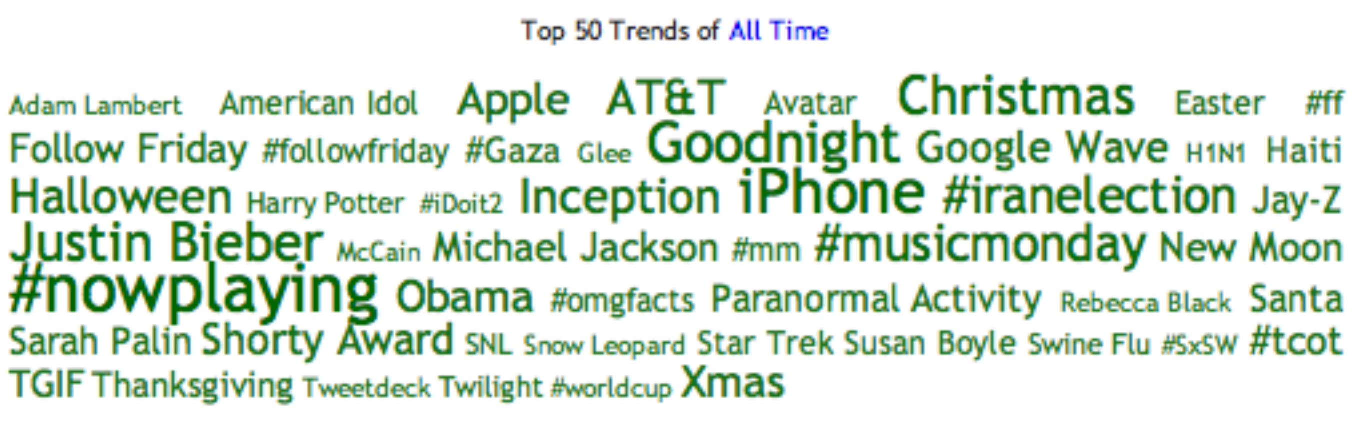
- Tag Clouds
  - Identify significant words
  - http://wordle.net

Flickr - All time most popular Photo Tags

Twitter Trending Topics



Twitter Trending Topics



Visualization Sample Links

- Collection of various forms of visualizations
  - http://www.smashingmagazine.com/2007/08/02/data-visualization-modern-approaches/
  - http://www.readwriteweb.com/archives/the_best_tools_for_visualization.php
  - http://www.visualcomplexity.com/vc/
  - http://social.cesweb.org/
  - IBM Many Eyes (http://www-958.ibm.com/software/data/cognos/manyeyes/)
  - Blogtrackers (http://blogtrackers.fulton.asu.edu/)

APPLICATIONS & RESEARCH TRENDS

Research Topics

- Influence
- Familiar Strangers
- Collective Wisdom
- Homophily
- Privacy in Social Media
- Collective Action

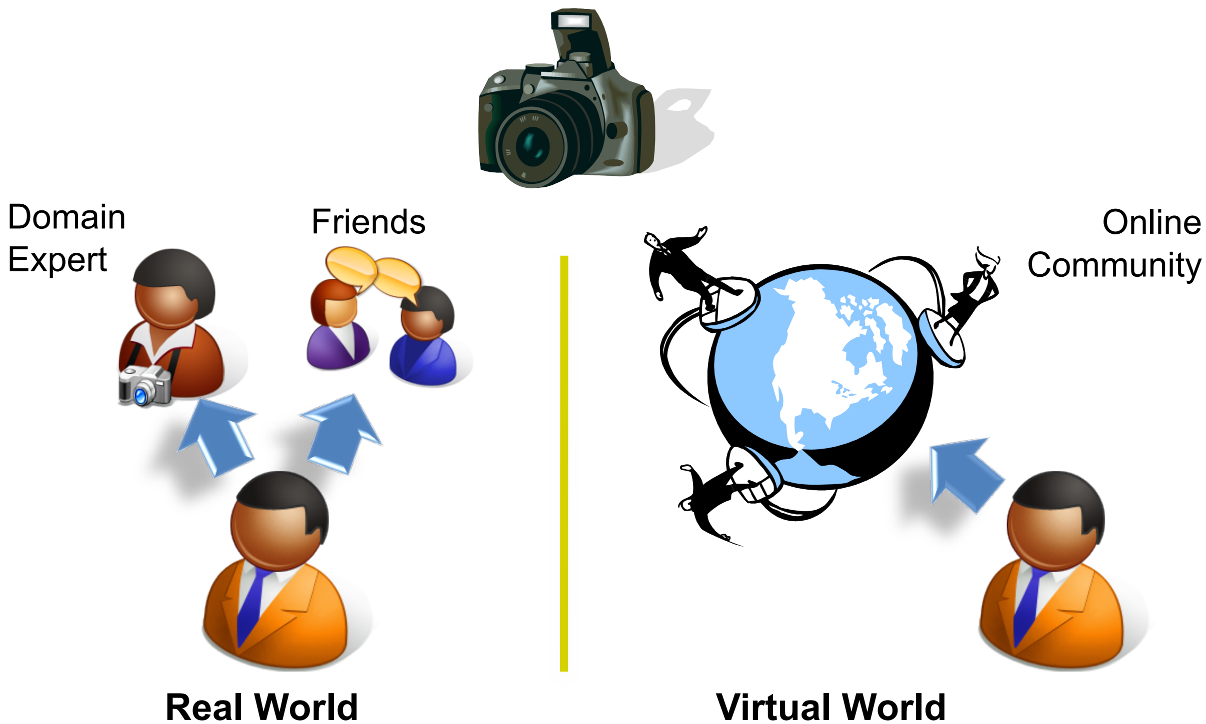
Identifying Influential Bloggers in a web community

Identifying the Influential Bloggers

as author at The 1st ACM International Conference on Web Search and Data Mining - WSDM 2008,
771 views

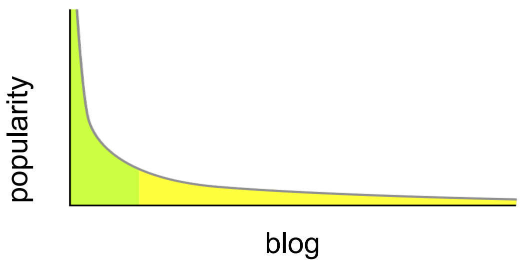
WSDM08 - http://videolectures.net/wsdm08_agarwal_iib/

Real and Virtual World



Influential Sites and Bloggers

- Power law distribution
- Short Head blogs
  – Influential sites
  – Search engines
  – Information Diffusion [Gruhl et al. 2004; Kempe et al. 2003; Richardson and Domingos 2002; Java et al. 2006]



- Long Tail blogs [Anderson 2006]
  – Inordinately many
  – Less popular
  – Cater to niche interests
- Extremely challenging to study all these blogs
- Influential bloggers as representatives

Influential Bloggers

- Inspired by the analogy between real-world and blog communities, we answer:
- Who are the influentials in Blogosphere?
- Can we find them?

Searching for the Influentials

- Active bloggers
  - Easy to define
  - Often listed at a blog site
  - Or, based on their blogging activity: submission rate
- How to define an influential blogger
  - Influential bloggers have influential posts
  - Subjective
  - Collectable statistics
  - How to use these statistics

?

Active Bloggers = Influential Bloggers

Active bloggers may not be influential

Influential bloggers may not be active

Intuitive Properties

- Social Gestures (statistics)
- **Recognition**: Citations (incoming links)
  - An influential blog post is recognized by many. The more influential the referring posts are, the more influential the referred post becomes.
- **Activity Generation**: Volume of discussion (comments)
  - Amount of discussion initiated by a blog post can be measured by the comments it receives. Large number of comments indicates that the blog post affects many such that they care to write comments, hence influential.
- **Novelty**: Referring to (outgoing links)
  - Novel ideas exert more influence. Large number of outlinks suggests that the blog post refers to several other blog posts, hence less novel.
- **Eloquence**: "goodness" of a blog post (length)
  - An influential is often eloquent. Given the informal nature of Blogosphere, there is no incentive for a blogger to write a lengthy piece that bores the readers. Hence, a long post often suggests some necessity of doing so.
- Influence Score = f(Social Gestures)

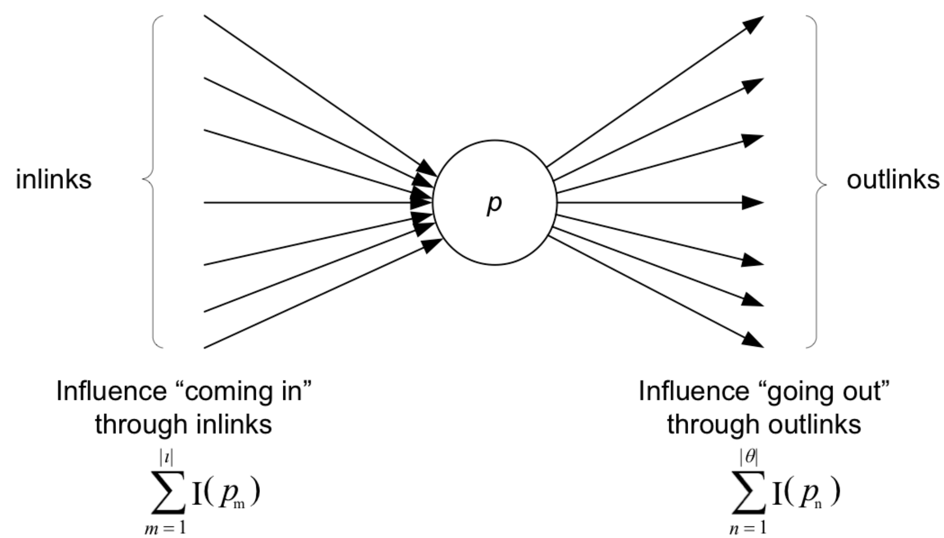
A Preliminary Model

$$InfluenceFlow(p) = w_{in} \sum_{m=1}^{|\iota|} I(p_m) - w_{out} \sum_{n=1}^{|\theta|} I(p_n)$$

$$I(p) \propto w_{comm} \gamma_p + InfluenceFlow(p)$$

$$I(p) = w(\lambda) \times (w_{comm} \gamma_p + InfluenceFlow(p))$$

$$iIndex(B) = \max(I(p_l))$$



The Unofficial Apple Weblog (TUAW)

TUAW The Unofficial Apple Weblog

iPhone Apple News App Store Mac 101 Macworld ~~iPhone~~

Filed under: **iPhone**

Filed under: Software, Internet, Internet Tools, iPhone

First Look: Analytics for iPhone

by Dave Caolo on Feb 5th 2009 at 8:00AM

Google Analytics is a popular and quite useful set of tools for monitoring a web site's traffic and performance. Set up is a snap and the reports are easy to read and flexible. You can create goals, monitor traffic and so on. What more could you want? On-the-go reports via your iPhone? All of your target statistics in your pocket? Oh, all right.

Earlier this week, Michael D Jensen of Inblosam LLC released Analytics App, which presents everything you'd ever want from Google Analytics on your iPhone. It is exhaustive.

When you first launch Analytics App, you're asked for your Google login (you must have a pre-existing Analytics account). From there, a list of all the sites you're monitoring appears. Click any one and view nearly 30 reports, including traffic, visitors, content ... even events tracking you've set up and your own customized reports. It's speedy over Wi-Fi and EDGE.

AT&T 8:00 PM
Reports www.thescificas... Date
Jan 5, 2009 - Feb 4, 2009
Visitors Overview
Visits - 1,080
112
58
5

Visits	1,080
Absolute Unique Visitors	762
Pageviews	1,994
Average Pageviews	1.85
Time on Site	00:02:18
Bounce Rate	66.02%

For example, Analytics App's traffic reports include referring sites, search engines, keywords, AdWords campaigns and more. Set the date range of any report to sort by day, week or month. The Dashboard provides an overview complete with easy-to-read graphs.

For $5.99US, this application is a keeper. Up-to-date stats from all of your sites, available nearly anywhere, makes our geeky little hearts go pitter-pat.

Gallery: Analytics App for iPhone

Tags: analytics app, AnalyticsApp, google, google analytics, GoogleAnalytics, iphone

Related Story Email this Share Comments (10)

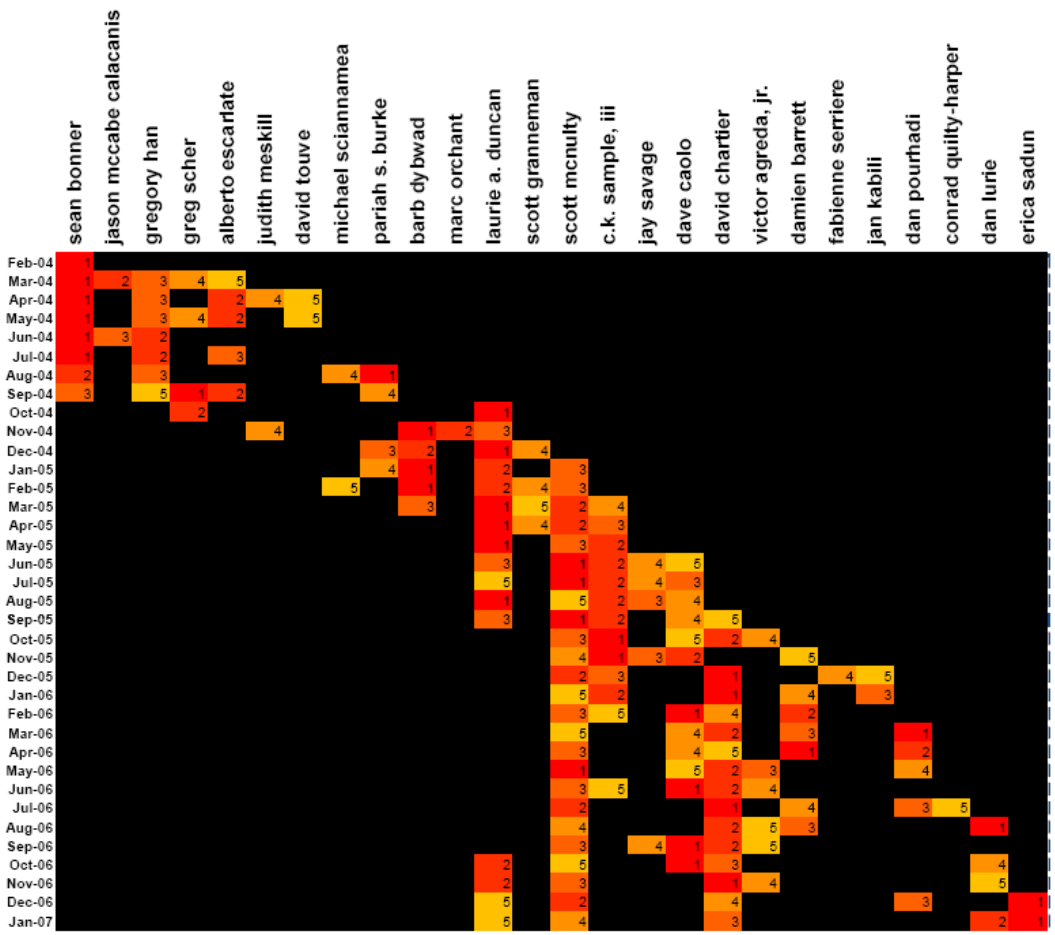
Active & Influential Bloggers

Top 5 TUAW Bloggers	Top 5 Influential Bloggers
Erica Sadun	*Erica Sadun*
Scott McNulty	Dan Lurie
Mat Lu	*David Chartier*
David Chartier	*Scott McNulty*
Michael Rose	Laurie A. Duncan

- Active and Influential Bloggers
- Inactive but Influential Bloggers
- Active but Non-influential Bloggers

- We don't consider "Inactive and Non-influential Bloggers", because they seldom submit blog posts. Moreover, they do not influence others.

Temporal Patterns

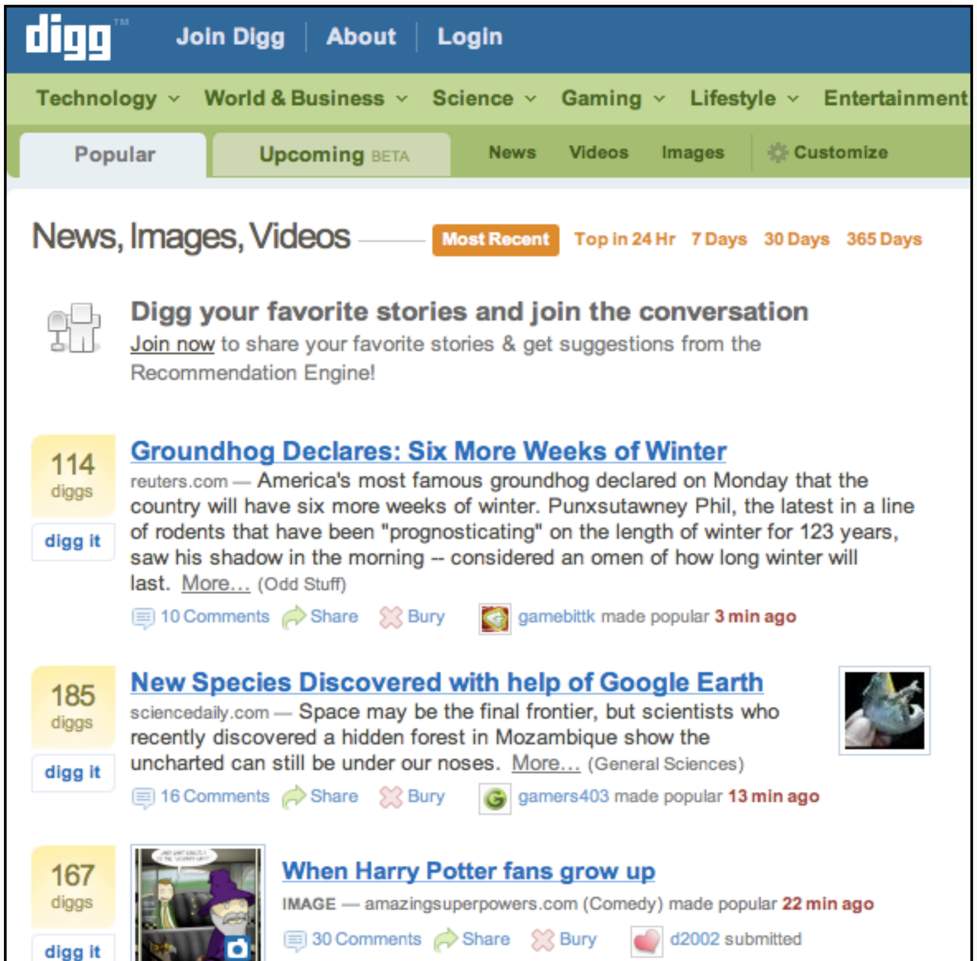


- Long term influential
- Average term influential
- Transient influential
- Burgeoning influential

Verification of the Model

- Challenges
  - No training and testing data
  - Absence of ground truth
  - How to do it?
- We use another Web 2.0 website, Digg as a reference point.
- "Digg is all about user powered content. Everything is submitted and voted on by the Digg community. Share, discover, bookmark, and promote stuff that's important to you!"
- The higher the digg score for a blog post is, the more it is liked.

Digg - Power of Web 2.0



Findings w.r.t. Digg

- Digg records top 100 blog posts obtained through Digg Web API.
- Top 5 influential and top 5 active bloggers were picked to construct 4 categories
- For each of the 4 categories of bloggers, we collect top 20 blog posts from our model and compare them with Digg top 100.

Bloggers	Active	Inactive
Influential	S1: 17	S2: 7
Non-influential	S3: 3	S4: 0/1

Bloggers	Active	Inactive
Influential	S1: 71	S2: 14
Non-influential	S3: 8	S4: 7

Bloggers	Active	Inactive
Influential	S1: 327	S2: 42
Non-influential	S3: 131	S4: 35

- Distribution of Digg top 100 and TUAW's 535 blog posts

Relative Importance of Parameters

- Compare top 20 blog posts from our model and Digg.
- Considered six months

	Jun 2007	May 2007	Apr 2007	Mar 2007	Feb 2007	Jan 2007
All-in	14	16	12	15	10	12
No Inlinks	3	4	3	3	1	0
No Comments	8	8	5	4	5	4
No Outlinks	11	8	5	4	4	7
No Blog post length	12	14	11	15	9	10

- Considered all configuration to study relative importance of each parameter.
- **Recognition (Inlinks) > Activity Generation (Comments) > Novelty (Outlinks) > Eloquence (Blog post length)**

Searching for Familiar Strangers

A Social Identity Approach to Identify Familiar Strangers in a Social Network

as author at The 2nd ICWSM 2009 – International AAAI Conference on Weblogs and Social Media,
48 views

ICWSM09 - http://videolectures.net/icwsm09_agarwal_siaifs/

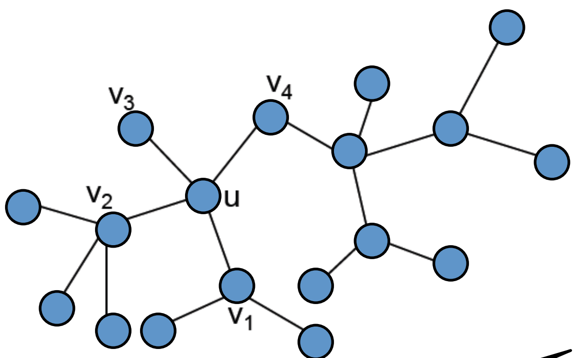
Who are Familiar Strangers

- Observe repeatedly, but do not know each other
- Real World
  - E.g., Individuals observe each other daily on a train
  - Discover the latent pattern: going to same workplace,
- Social media
  - You are defined by what you share…
  - Have similar interests (Movie, Games, Technology, Politics)
  - Not in each others social network
- Identify "Good strangers" in hostile situations

Aggregating Familiar Strangers

- Together they form a critical mass
  - understanding of one individual gives a sensible and representative glimpse to others
  - better customization and services (e.g., personalization and recommendation)
  - nuances among them present new business opportunities
  - predictive modeling and trend analysis

An Example



u: Given individual

C_u: {v_1,v_2,v_3,v_4}
A_u: {Exercise, History, Recreation}
A_{v1}: {Internet, News}
A_{v2}: {Blogging, Internet}
A_{v3}: {Blogging, Internet, Technology}
A_{v4}: {Recreation, Travel}

Egocentric network view

Find T_u, given γ: Sports={Exercise, Recreation}

Social Identity Approach

- Social Identity: ability to cluster contacts into meaningful groups [Tajfel, H. 1978]
- Propagate the search through relevant clusters of contacts
- Prunes the search space
- Desiderata
  - Small-world assumption [Watts and Strogatz 1998]
    - Power law degree distribution:
    - High clustering coefficient:
    - Short average path length:

$$f(x) \propto ax^{-\gamma}$$

$$\kappa_v = \frac{2E_v}{|C_v|\left(|C_v| - 1\right)}$$

$$l_G = \frac{1}{n(n-1)} \sum_{\substack{i,j \\ i \neq j}} d(v_i, v_j)$$

Social Identity Construction

- Offline clustering of contacts
- Contacts represented by
  - Tag vector
  - Content vector
- LSA transformation to concept vectors [Deerwester et al. 1990]

$$X_{tag} = U_{tag} \Sigma_{tag} V_{tag}^{T} \qquad X_{con} = U_{con} \Sigma_{con} V_{con}^{T}$$

- S_{tag}: Pairwise cosine similarity between row vectors of V_{tag}
- S_{con}: Pairwise cosine similarity between row vectors of V_{con}
- $S = \alpha S_{tag} + (1-\alpha) S_{con}$
- k-means clustering

Experiments

- Ground Truth - Global network view
  - Steiner tree based approach [Du and Hu 2008]
  - Lower bound on search space
- Compare with
  - Exhaustive approach
  - Random approach
- Datasets:
  - Blogcatalog (~24K nodes)
  - DBLP (~35K nodes)

Alternative Approaches

- Exhaustive Approach
  - Search all the contacts
  - 100% accuracy
  - Exponential search cost: $\sum_{k=1}^{h} d^k$
- Random Approach
  - Fraction of contacts (σ) propagate the search
  - σ = 1 corresponds to Exhaustive approach

Results

- Blogcatalog

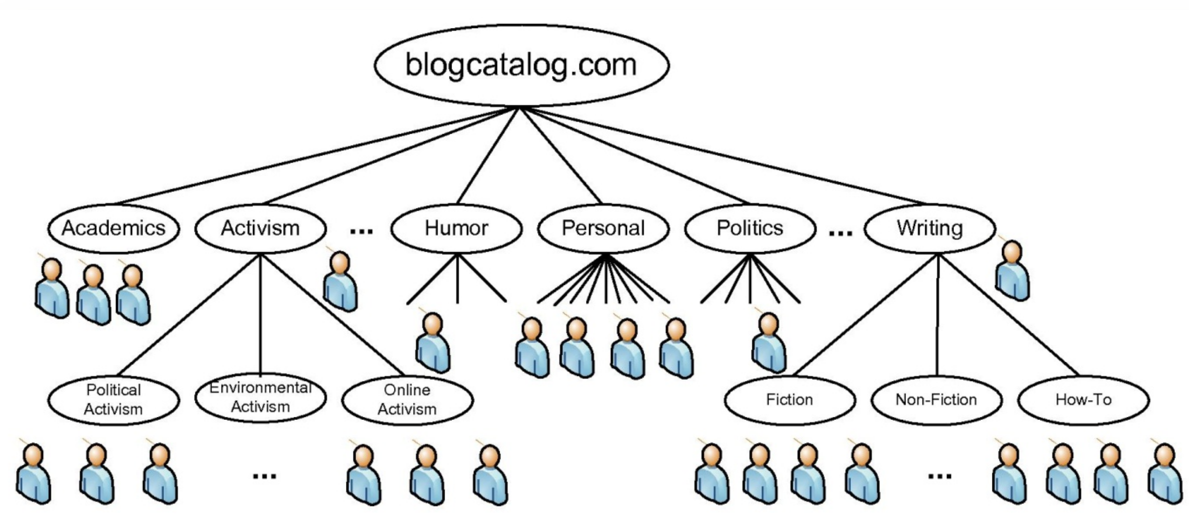
Approach (E)	*Accuracy (%)*	*Search performance (edge traversals)*
Steiner Tree	100%	3,565 ± 23
Exhaustive	100%	4,531,967 ± 944
Random	1.0283% ± 0.928	1,823 ± 43
Social Identity	79.2908% ± 3.008	6,032 ± 46

- DBLP

Approach (E)	*Accuracy (%)*	*Search performance (edge traversals)*
Steiner Tree	100%	4,752 ± 30
Exhaustive	100%	909,543 ± 403
Random	2.304% ± 0.355	58 ± 12
Social Identity	91.349% ± 2.107	12,182 ± 68

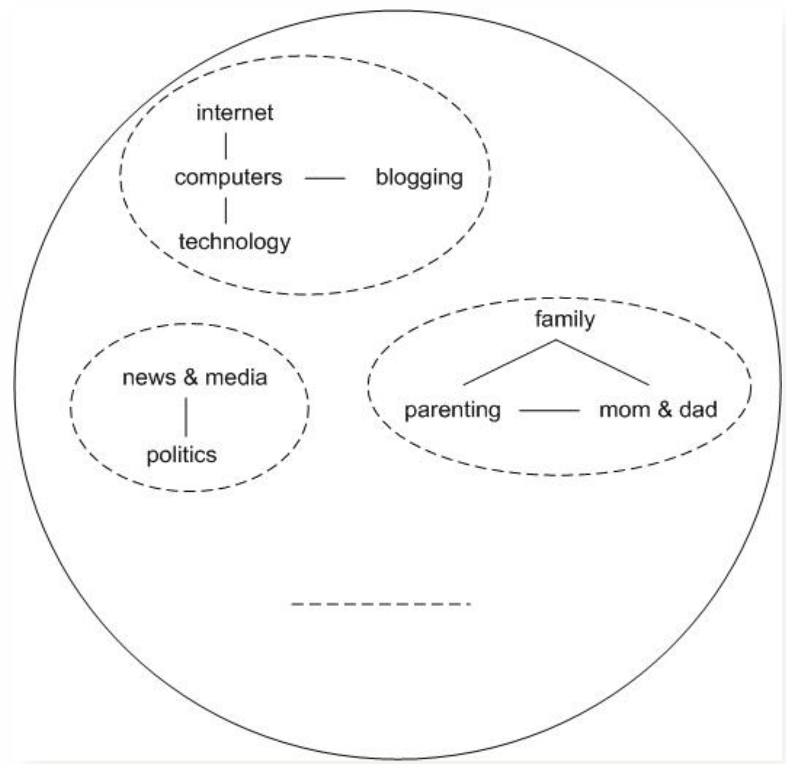
Leveraging Collective Wisdom (ICDM'09)

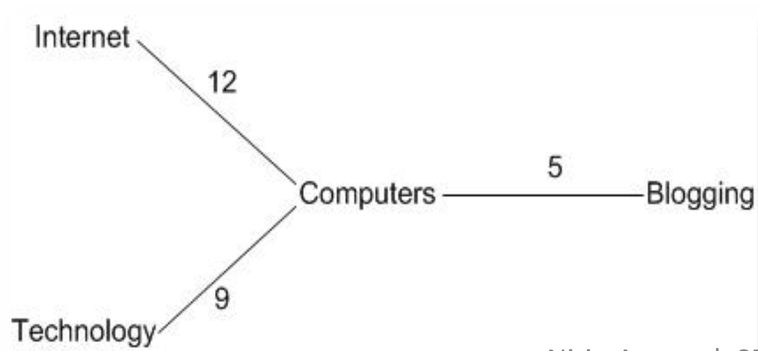
- Bloggers
  - Submit blogs, blog posts
  - Assign blog tags, category descriptors
- 56 categories in hierarchical fashion at blogcatalog.com



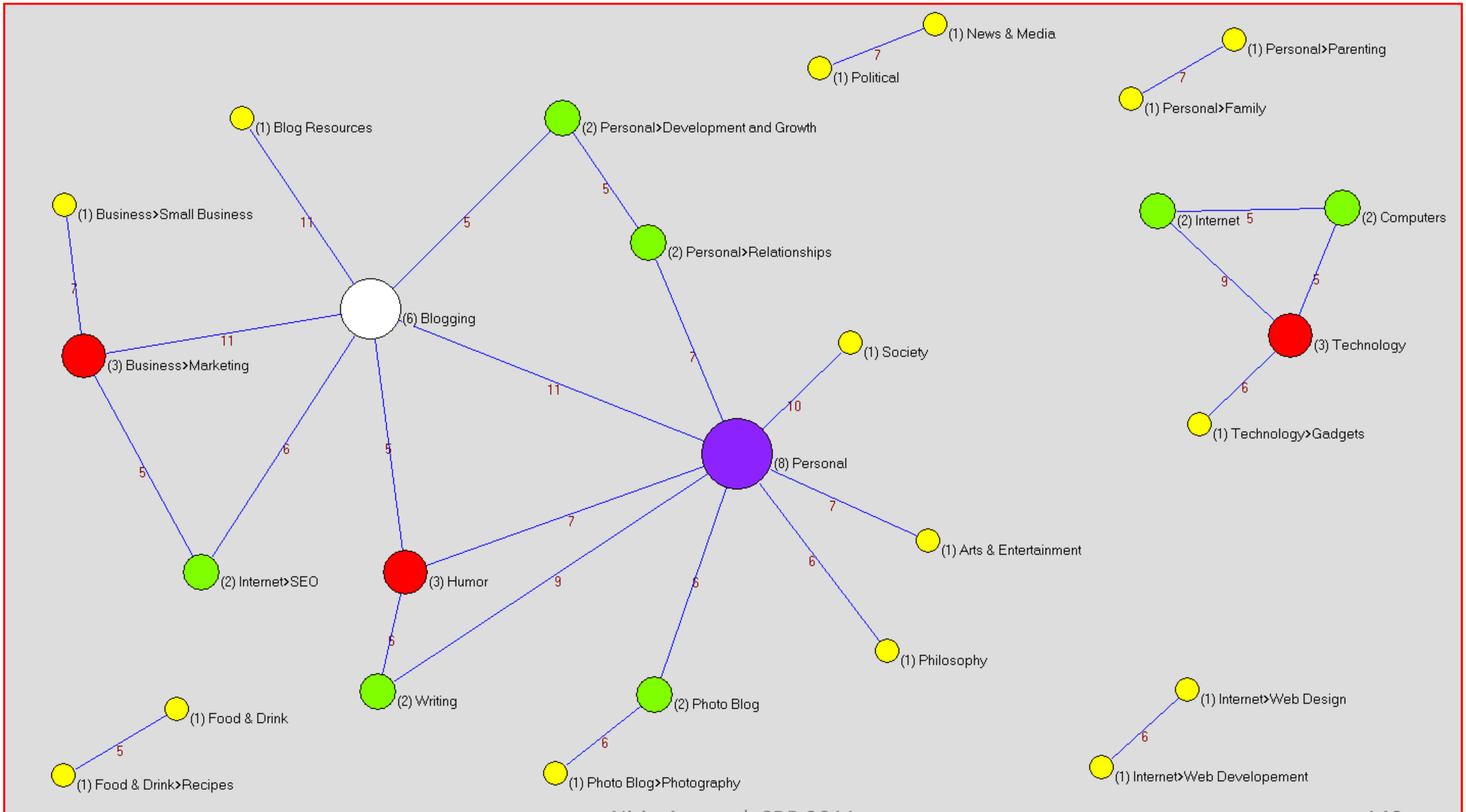
Category Relation Graph

- Connects categories that are simultaneously used by the bloggers
- Weights on the edges (Link Strength) denote the semantic relatedness of categories
- Link strength values are normalized between 0 and 1



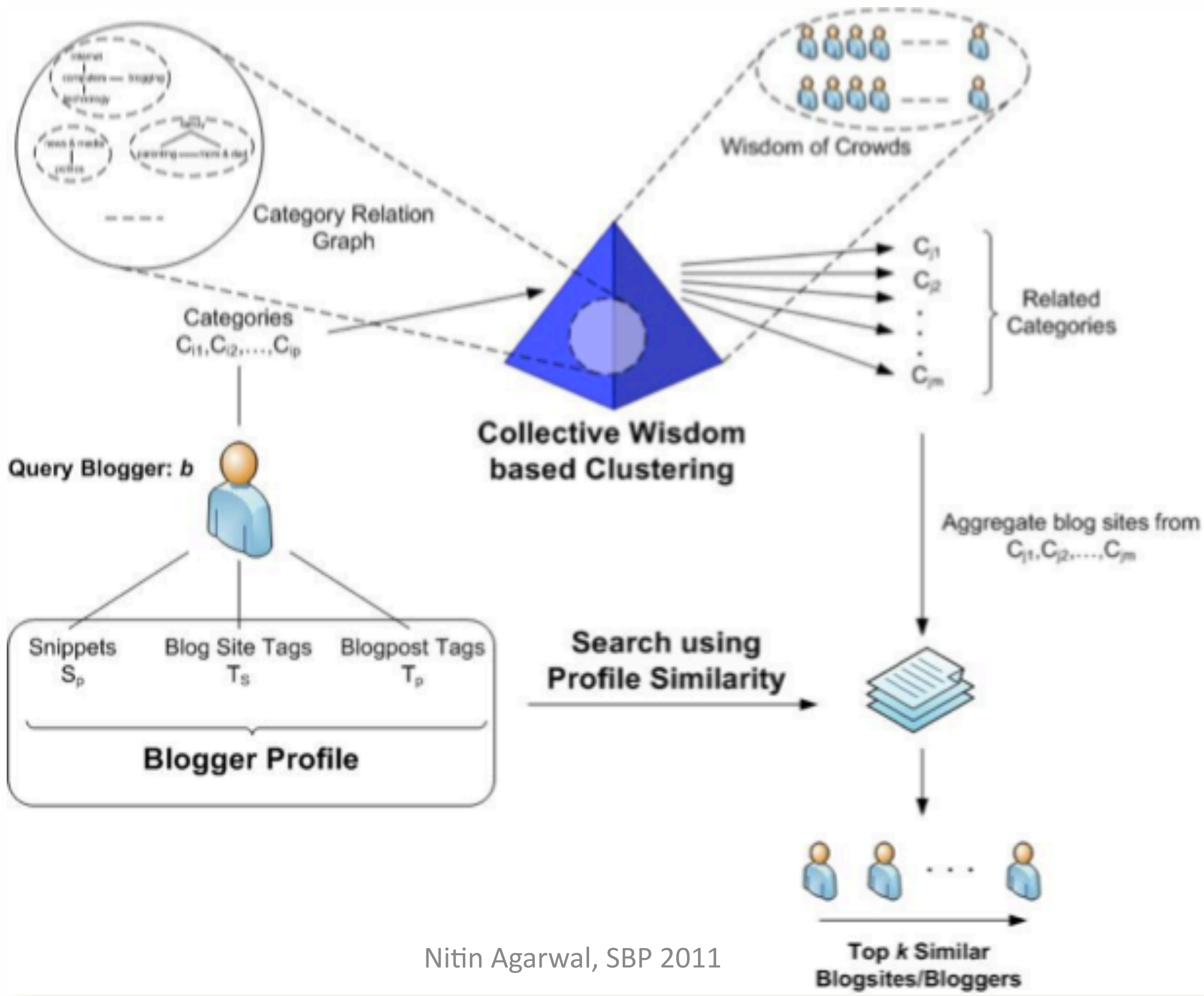


Visualization



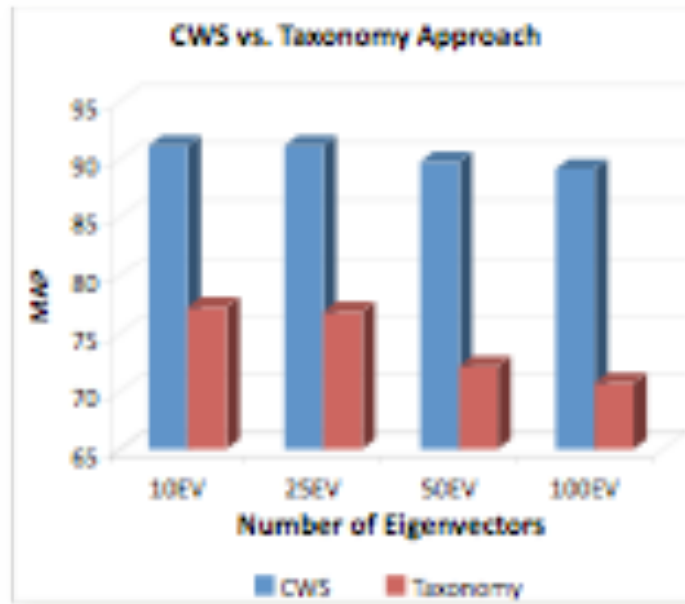
Pajek was used to generate this visualization

Searching for Similar Bloggers using Collective Wisdom

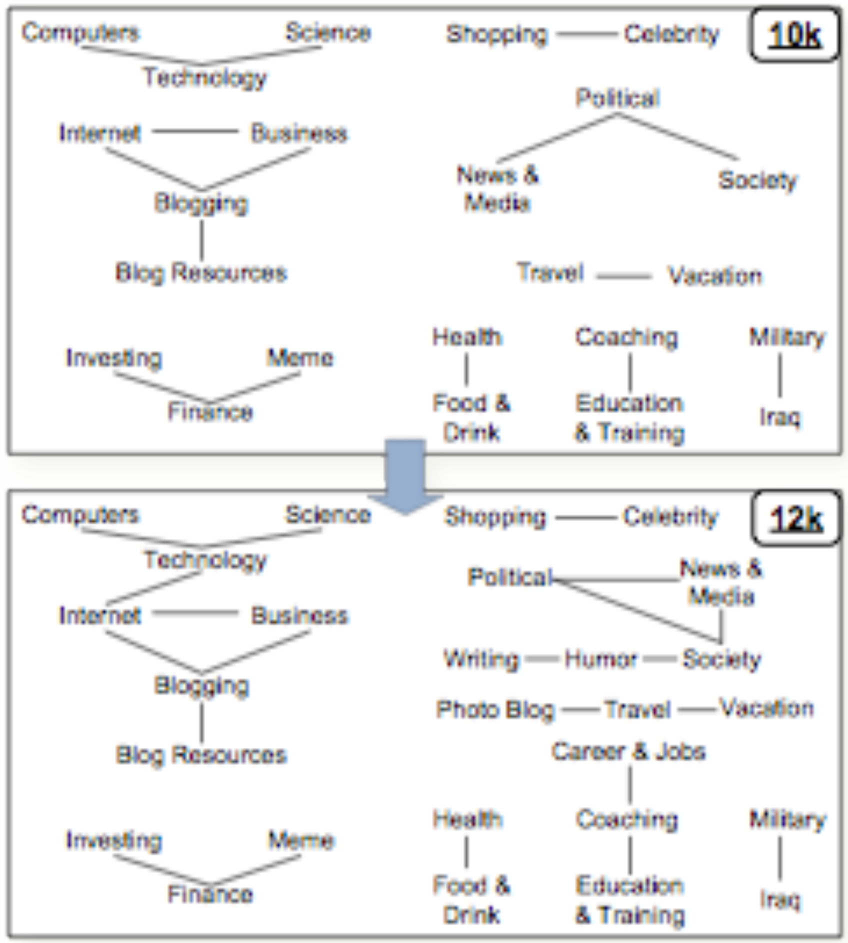


Collective Wisdom vs. Taxonomy based Search

- Collective Wisdom Search (CWS)
  - Accuracy: 91.164%
  - Search Space Reduction: 28.723% (with respect to exhaustive search)
- Taxonomy Search
  - Accuracy: 77.254%
  - Search Space Reduction: 42.084% (with respect to exhaustive search)
- If the skewed category distribution is neglected Search space reduction for CWS increases to 51.526%



Dynamics of Collective Wisdom



As tagging behavior changes over time, we observe changes in the category relation graph

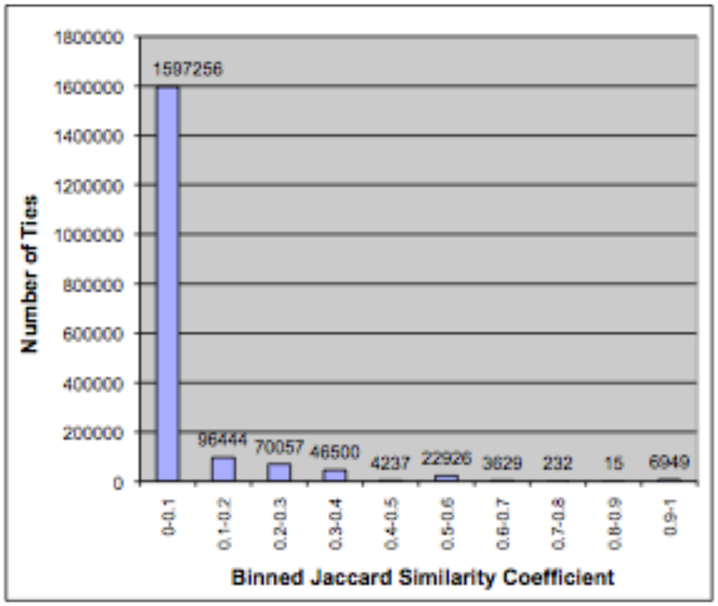
Homophily (KDD'10, WI'10)

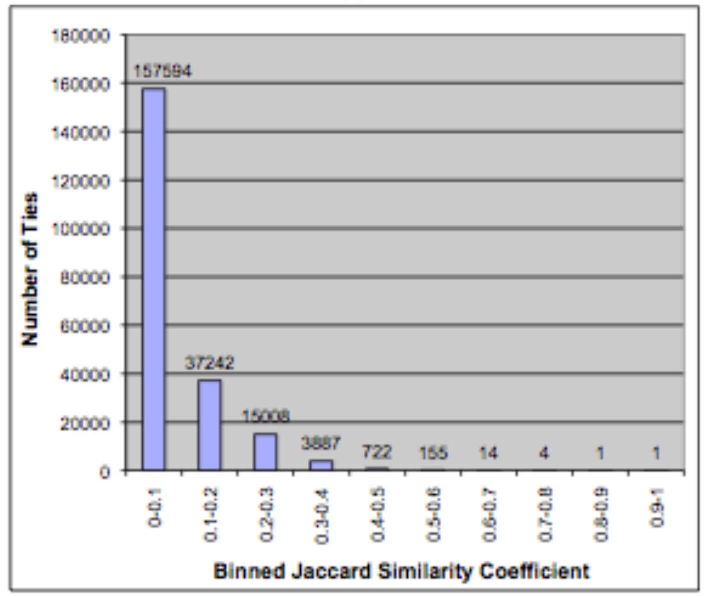
- Birds of a feather flock together
- Homophily - similar individuals are assumed to associate with each other more often than others

Physical World	Online/Virtual World
Sociodemographic dimensions such as age, gender, education, social status used to study homophily.	Sociodemographic dimensions are often not available or could not be trusted.
Physical locality such as geographical proximity and organizational locality such as workplace, schools play significant role in governing new ties.	Interactions between individuals span all geographical barriers across different timezones. Geographical or organizational proximity do not govern construction of ties.
User interests, opinions, thoughts, perspectives, and preferences were often ignored in studies conducted in physical world scenario.	Individuals on social media are defined by what they write/share. Interests, opinions, thoughts, perspectives, and preferences are the significant dimensions that could govern new ties.
Construction of new ties in physical world are often regulated by social status or class.	Construction of ties in virtual world are beyond social status and class.
Studies conducted in physical world were often limited to a particular geographical area constraining the scale of the study.	Millions of individuals could be easily studied in virtual world as compared to physical world. This makes the results much more conclusive and generalizable.

- Given the differences between real-world and online social media, does homophily exist in online social media?

Dyadic Relations



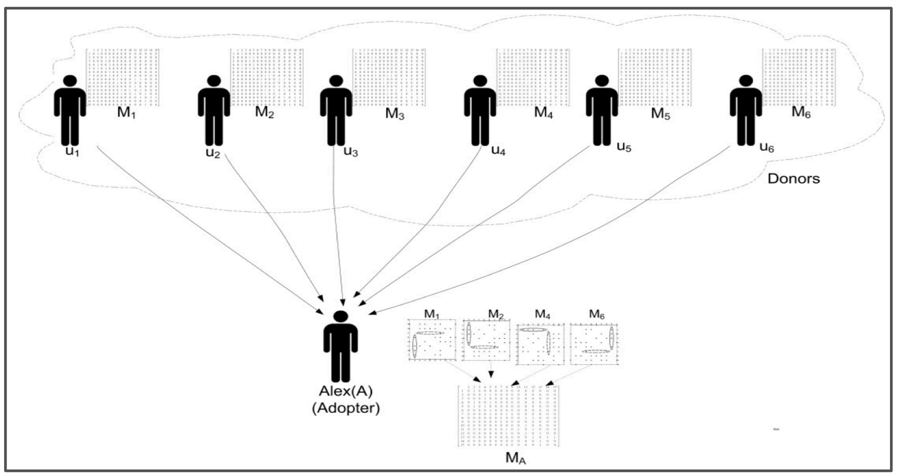


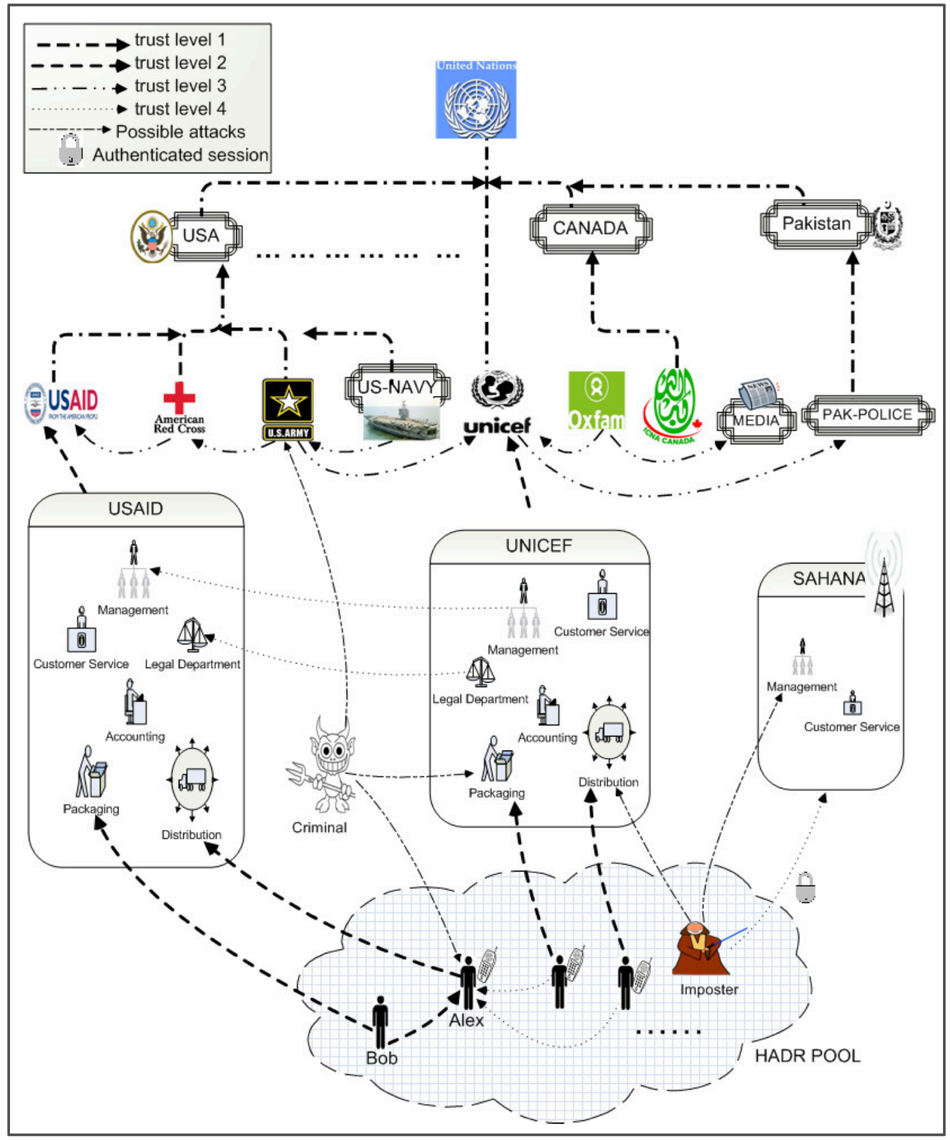
	Blogcatalog	Last.fm
0 interests in common	84.0609%	23.2490%
1 or more interests in common	15.9390%	76.7509%

Privacy

[NSF Science360, US News, UALR News, PhysOrg, Dr Dobbs Journal]

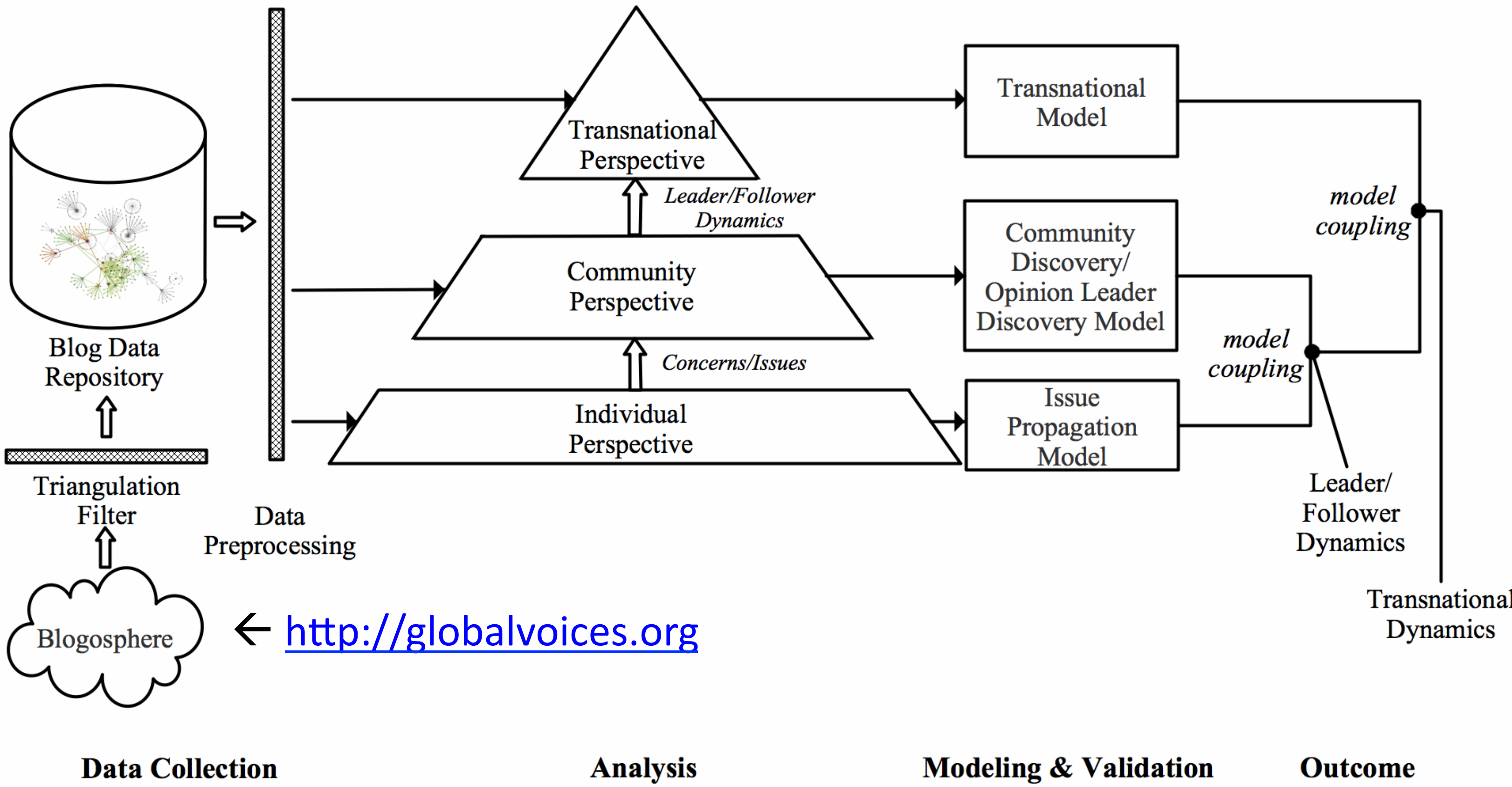
- Context Based Privacy Model
- Best paper award at IEEE international conference on Privacy, Security, Risk, and Trust (PASSAT 2010)
- Collective model (HA/DR) to respect/evaluate trust and privacy





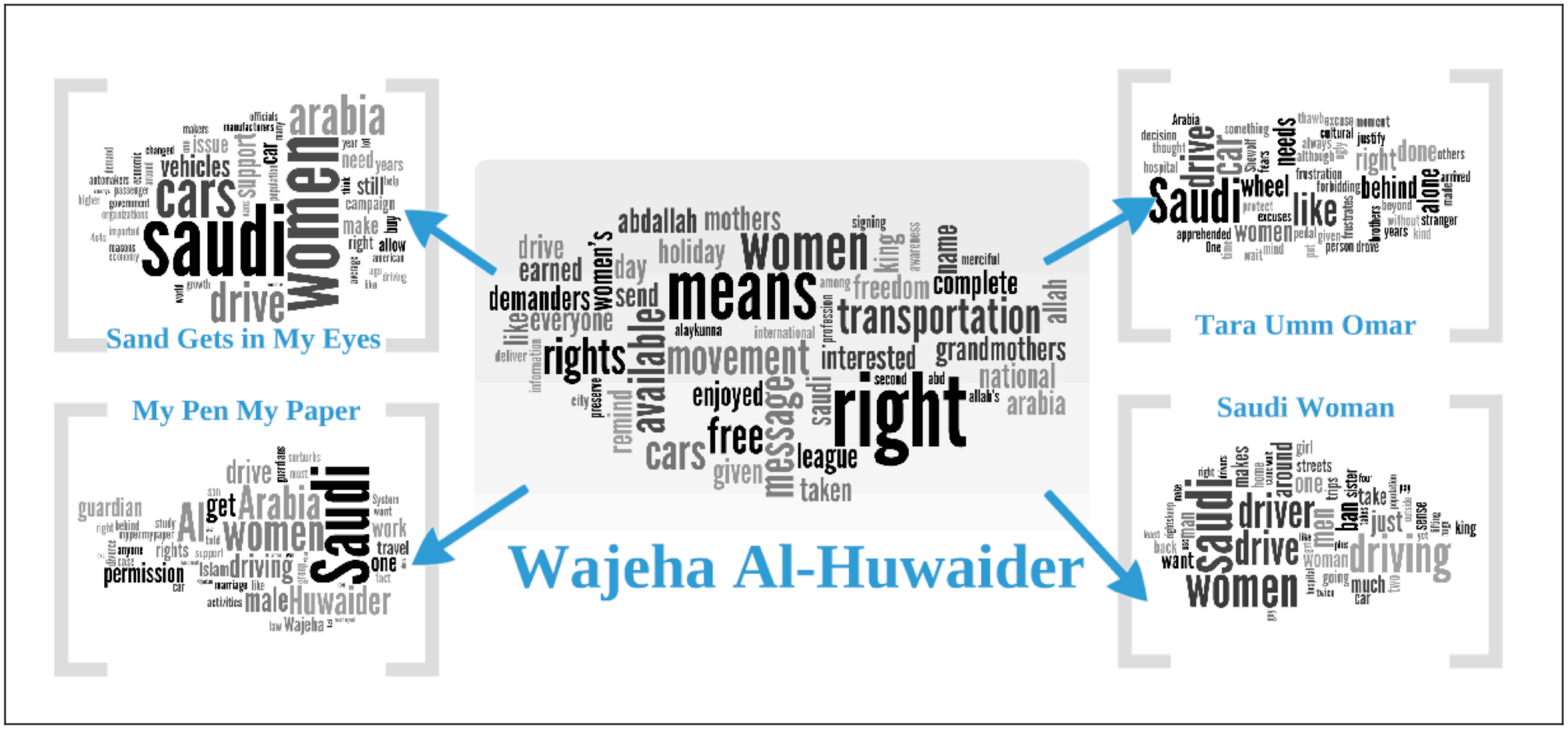
Collective Action (NDT'11, ECIS'11)

Egyptian, Tunisian uprisings → convulsions of revolution
Unorganized yet strategic communications → "unorganized organizations"
Attempt to analyze this phenomenon using the social media

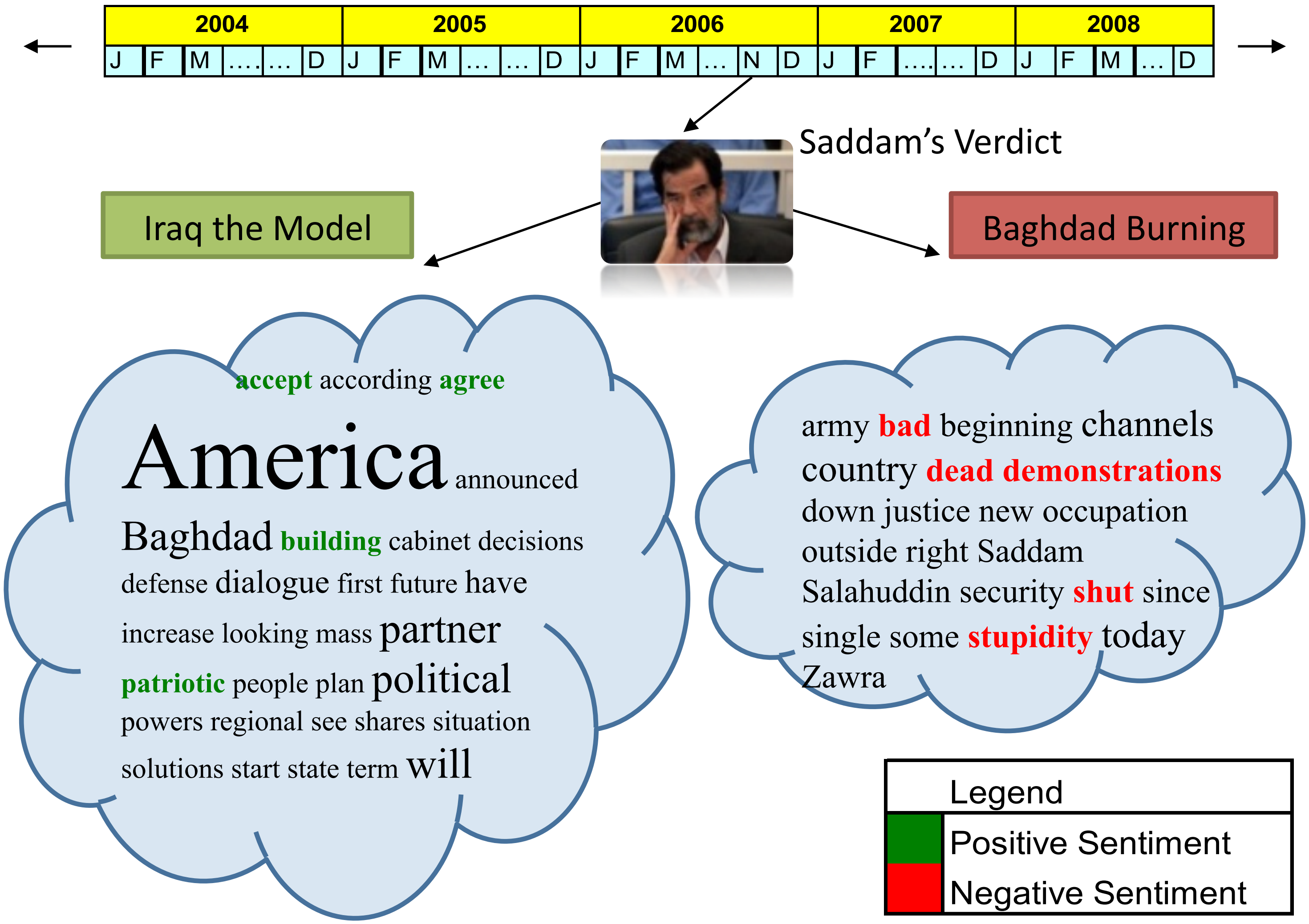


← http://globalvoices.org

Individual Perspective

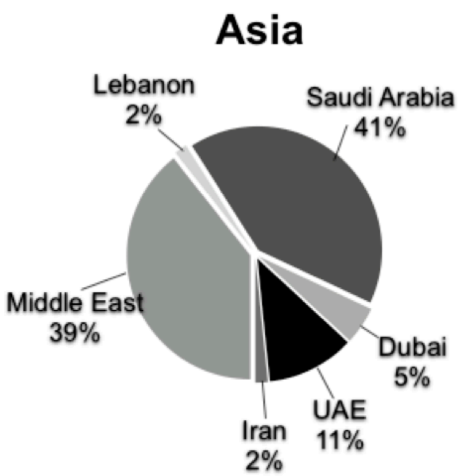


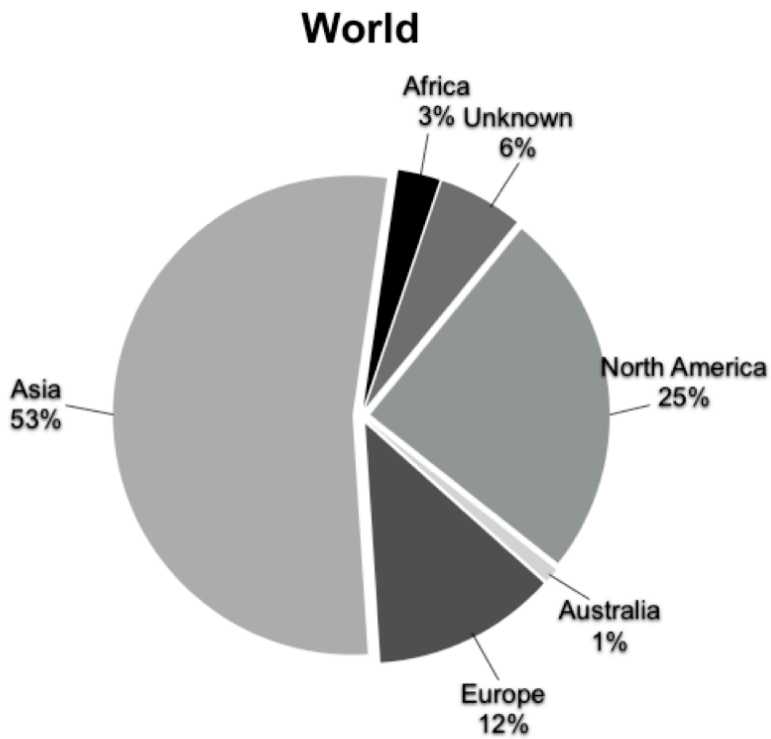
Community Perspective

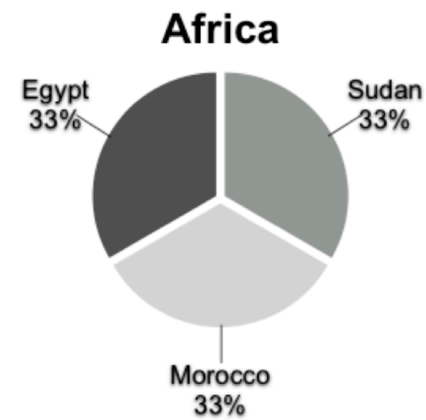


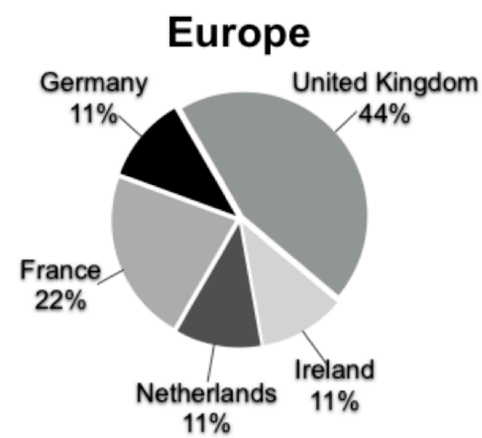
Understanding group interactions (ICCCD'08)

Transnational Perspective









Revolution 2.0

Social media and political changes in Egypt and beyond

Online activism in the Middle East did not begin in Tahrir Square on January 25, but has been evolving for many years. In this lecture, Merlyna Lim will chronicle how the Internet, including social media, facilitated the emergence of new networks of opposition to the ruling regime in Egypt, and how such networks and their converging narratives were translated into coordinated mass actions that led to a relatively peaceful overthrow of a dictatorship.

Lecture & Discussion
Tuesday, March 29
6:30PM

by Merlyna Lim, PhD

followed by discussion with
Chad Haines, PhD

ASU Tempe Campus
Coor Building, L1-74

Webcast
http://www.ustream.tv/channel/cspo

Merlyna Lim, PhD
Assistant Professor, School of Social Transformation and Consortium for Science, Policy & Outcomes

Merlyna Lim has studied the mutual shaping of society and technological systems – including the Internet and social media – for more than a decade, with a particular focus on social media activism in the Middle East since 2007. She has published extensively about the politics of information technology in Muslim societies.

Co-Presented by Arizona State University's
Consortium for Science, Policy & Outcomes
Center for the Study of Religion and Conflict

THE CENTER FOR THE STUDY OF RELIGION AND CONFLICT

Chad Haines, PhD
Research Fellow and Lecturer, Center for the Study of Religion and Conflict

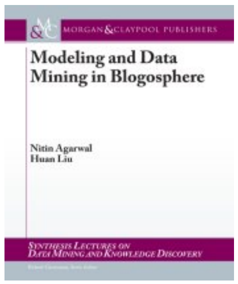
Chad Haines is a cultural anthropologist whose research engages the complex ways postcoloniality and globalization reshape the Muslim world. He formerly was an assistant professor of anthropology at American University in Cairo.

Acknowledgments

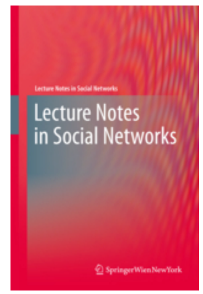
- Huan Liu, ASU; Yi Chen, ASU; Philip S. Yu, UIC; John Salerno, AFRL; Mark Woodward, Center for Religious Studies at ASU; Sun-ki Chai, University of Hawaii; Arunabha Sen, ASU; Xiaowei Xu, UALR; Rolf T. Wigand, UALR; Merlyna Lim, ASU; Thomas A. Schweiger, Acxiom; Hemant Joshi, Acxiom

- AFOSR and ONR grants

Additional Information



New book on Modeling and Data Mining in Blogosphere
Over 800 downloads, highest on publisher's website
2010

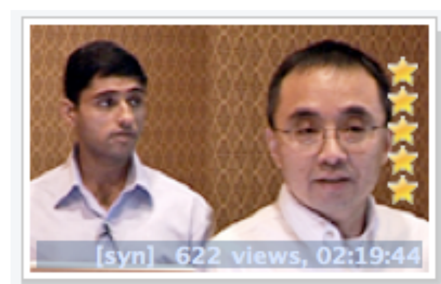


Edited book in Lecture Notes in Social Networks series on Online Collective Action: Dynamics of the Crowd in Social Media
2012

New book on
Social Computing in Blogosphere: Challenges, Methodologies, and Opportunities
2011

Special Issue on Social Computational Systems, Elsevier Journal of Computational Science, to appear in 2011, With Xiaowei Xu

Special Issue on Social Computing in Blogosphere in
IEEE Internet Computing Magazine
Co-Editors: Huan Liu, Philip S. Yu, and Torsten Suel.
Issue: March-April 2010.



KDD 2008 Tutorial on Research Opportunities and Challenges in Blogosphere