

Social Computing and Its Application in Query Suggestion

Irwin King

king@cse.cuhk.edu.hk

<http://www.cse.cuhk.edu.hk/~king>

Department of Computer Science & Engineering
The Chinese University of Hong Kong



Billionaires' Shuffle

2007



2008



Facebook in 2004.02

2008
at **23** and **\$1.5** billion later...



Alexa as of Nov. 2008	USA	CHINA	Global
1	Google	Baidu	Yahoo
2	Yahoo	QQ	Google
3	Myspace	Sina	YouTube
4	YouTube	Google.cn	Windows Live
5	Facebook	Taobao	Facebook
6	Windows Live	163	MSN
7	MSN	Yahoo	Myspace
8	Wikipedia	Google	Wikipedia
9	EBay	Sohu	Blogger
10	AOL	Youku	Yahoo.jp



What's On the Menu?

- Web 2.0 and Social X
- Social Computing
- Some Interesting Problems
 - Collaborative Filtering
 - Query Suggestion



What's On the Menu?

- Web 2.0 and Social X
- Social Computing
- Some Interesting Problems
 - Collaborative Filtering
 - Query Suggestion



Web 2.0

- Web as a medium vs. **Web as a platform**
- Read-Only Web vs. **Read-and-Write Web**
- Static vs. **Dynamic**
- Restrictive vs. **Freedom & Empowerment**
- Technology-centric vs. **User-centric**
- Limited vs. **Rich User Experience**
- Individualistic vs. **Group/Collective Behavior**
- Consumer vs. **Producer**
- Transactional vs. **Relational**
- Top-down vs. **Bottom-up**
- People-to-Machine vs. **People-to-People**
- Search & browse vs. **Publish & Subscribe**
- Closed application vs. **Service-oriented Services**
- Functionality vs. **Utility**
- Data vs. **Value**



Social Networking

The screenshot shows the Facebook profile of Irwin King. The profile includes a profile picture, a cover photo, and a bio. The bio states: "What are you doing right now?" and lists his networks as CUHK Faculty, sex as Male, hometown as Taipei, Taiwan, and religious views as Christian. The Mini-Feed shows several updates, including a new address at the Department of Computer Science and Engineering, The Chinese University of Hong Kong, and several new friendships with Chi Chung Chan, Tom Hung, and David Shepherd. The Information section lists his contact info, current address, and website. The Education and Work section lists his grad schools: University Of Southern California '88 M.Sc., Computer Science; University Of Southern California '03 Ph.D., Computer Science.

The screenshot shows the Myspace.com profile of drmanhattan. The profile includes a profile picture, a cover photo, and a bio. The bio states: "Like all good first albums should, the debut from drmanhattan is filled with throbbing punk melodies and the kind of lyrics you'll find yourself singing along to on your second listen. Listen to the album now, before it hits stores 3/11, exclusively on MySpace." The profile also features a "Cool New Videos" section with several video thumbnails, a "MySpace Music" section with a "MySpace Secret Stand Up Presents... Aisha Tyler in Boston!" advertisement, and a "Cool New People" section with several profile pictures. The profile also includes a "Member Login" section and a "Find Your Friends on MySpace" section.



Social Search

- Social Search Engine
- Leveraging your social networks for searching

eurekasterswicki login | sign up

build new swicki | swicki directory | about swickis | about eurekaster

Search and vote for your faves

swicki search

a custom search portal around the topic of your choice powered by your community

Build a swicki!

A swicki is a custom social search portal on the topic of your choice. With every search, vote and click, your swicki generates more relevant results and turns into a valuable asset for you and your community. Take a tour to find out more about how swickis work.

- Choose from text, multimedia or video content
- Customize the swicki widget look and feel
- Share your swicki widget with your community

Build a swicki

New! Even fresher swickis with RSS and Autodetect. [Learn More.](#)

Eurekaster news

Now out of beta!

- Come join the network for swicki builders
- Swicki Users Go Green
- CEO Speaking at SES New York

Get swicki illustrated
For the latest news and trends in social search, subscribe now.

Browse the directory

Try searching one of over 100,000 swickis already created, or grab one to add to your site or blog.

Recently created

- askforkids
- e-learning et didactique ...
- denver news
- home repairs any gal can ...
- creative ideas for green ...
- easy woodworking projects ...

More >

Top swickis

- techcrunch
- borr2ikes
- popular science
- readrteweb
- lockergnome
- neopets
- larkswicki

More >

DIY: home improvement swicki showcase

- Home Repairs Any Gal Can Do
- Make Yourself a Man Pad
- Making Room for Baby
- Creative Ideas for Green Home Improvement

Computers

- dot net search engl...
- php resource search
- rails on ruby
- software factories
- web 2.0 workgroup

More >

Business

- adblogging
- alternative search ...
- bubblegeneration - ...
- contextual advertis...
- digging into search
- freelance tipster
- green building reso...

More >

Home

- about color for hom...
- gardening and plant...
- home improvement se...
- homemade baby food ...
- homemaking
- salmon

More >

Regional

- amazon river
- atlanta business se...
- atlanta home and ga...
- berkeley public lib...
- pittsburgh news
- pittsburgh wedding ...
- ski tahoe

More >

delver:: liad agmon edit

My Profile | My Network

Your friends are the best source of information!
Look for information, media and people within your network

(Go)

Noa Rabiner
Noa Rabiner is connected to you directly

- This is me!
- I know this person
- Add as Connection
- Send Message



Social Bookmarking

The screenshot shows the del.icio.us website interface. At the top left is the logo and tagline "social bookmarking". Navigation links include "login", "register", and "help". A search bar is located at the top right. The main content area features a "hotlist" section titled "what's hot right now on del.icio.us" with a "HOT NOW" sub-header. It lists several bookmarked items with their titles, authors, and the number of people who bookmarked them. For example, "Video: Twitter in Plain English" has 130 people bookmarking it. To the right of the hotlist is a "Tags" section with a definition and a "tags to watch" list including "illustration", "family", "living", "cool", and "itunes".

del.icio.us
social bookmarking

del.icio.us search
login | register | help

» all your bookmarks in one place
» bookmark things for yourself and friends
» check out what other people are bookmarking

learn more... » get started «

hotlist what's hot right now on del.icio.us

HOT NOW see also: popular | recent

Video: Twitter in Plain English | Common Craft - Explanations In Plain English 130 people
save this
first posted by jtyerse twitter video howto commoncraft web2.0 tags

Home | NotchUp Beta save this 212 people
first posted by sokrates_af jobs interview career search job tags

PrimeTimeRewind - The TV Cube save this 145 people
first posted by david.rothman tv video streaming television media tags

The Simple Dollar » Planning a Kitchen Garden save this 133 people
first posted by lantzilla gardening food garden cooking vegetables tags

Office Live Workspace vs Google Docs: Feature-by-Feature Comparison - ReadWriteWeb save this 135 people
first posted by gariig microsoft google office google_docs live tags

Tags
A tag is simply a word you use to describe a bookmark. Unlike folders, you make up tags when you need them and you can use as many as you like. The result is a better way to organize your bookmarks and a great way to discover interesting things on the Web.
learn more...

tags to watch more ...

illustration
karenklassenillustration
Dave Devries's Monster Engine
current work

family
Cozi
Comeeko - Creating comic strips from your photos
Let's Have More Teen Pregnancy

living
Home - tiny living
Eartheasy homepage
The Simple Dollar » Nourishment on a Desperate Income

cool
Browse Goods
rssWheel
Laptop Stand By LapDawg - A Revolutionary, Ergonomic Laptop Holder

itunes



Social Media

The screenshot shows the YouTube homepage with the following elements:

- Header:** YouTube logo, "Broadcast Yourself™", navigation tabs (Home, Videos, Channels, Community), search bar, and "Upload" button.
- Videos being watched right now...:** A row of five video thumbnails.
- Promoted Videos:** Four video thumbnails with titles like "Think Again Awards" and "第14届十大电视广告颁奖典礼 - 播出...".
- Featured Videos:** A list of featured videos with titles and view counts:
 - David Sedaris delivers a pizza:** From *weeknight*, Views: 11,313, 5 stars, 01:01.
 - Erbert and Gerbert's Candle Cannon:** From *candlecannon*, Views: 109,029, 5 stars, 02:34.
 - Girl's Night Out:** From *girlsnite*, Views: 169,435, 5 stars, 03:49.
 - Lionel Neykov - Freeze My Senses:** From *LionelNeykov*, Views: 150,758, 5 stars, 03:35.
- What's New:** A yellow box containing:
 - YouTube Mobile:** "New! Watch ALL YouTube videos on your mobile device".
 - Warp!** "Visually fly through YouTube videos in the Fullscreen player".
 - RSS Feeds:** "Click on the 'RSS this page' link to get fresh videos delivered".
 - SXSW on YouTube:** "For the next week and a half, the SXSW festival is taking over Austin, Texas, to celebrate music, film and all things interactive. [Read more in our Blog](#)".

The screenshot shows the Flickr homepage with the following elements:

- Header:** Flickr logo, "Create Your Account" button, and "Sign In" link.
- Main Content:** A large photo of a small plant growing in a crack in the pavement. Text reads: "Share your photos. Watch the world." Below is a search bar and a "SEARCH" button.
- Statistics:** "3,802 photos uploaded in the last minute · 558,832 photos tagged with urban · 2.2 million photos uploaded this month · [Take the tour](#)".
- Navigation:** Four icons with labels: "Share & stay in touch", "Upload & organize", "Make stuff!", and "Explore...".
- Footer:** "Take the Tour" button and a link to "Explore Flickr Blog, the World Map, Camera Finder or interesting photos from the last 7 days".

The screenshot shows the Second Life homepage with the following elements:

- Header:** "SECOND LIFE" logo, "Your World. Your Imagination.", and navigation links (What is Second Life?, Showcase, Community, Blog, Support).
- Main Content:** A large image of a man and a woman flying in a virtual world. Text reads: "Get Started! Membership is FREE! Second Life is an online, 3D virtual world imagined and created entirely by its Residents." Below is a link: "Discover a whole new world of friends, fashion, music, videos and fun! Explore the best of Second Life >>".
- Footer:** "Your Organization in Second Life! Find out why your business, school or nonprofit organization should get its own virtual world presence." Below is a "Visit Second Life Now!" button and several smaller images related to the virtual world.

Social News/Mash Up

The screenshot displays a social news/mashup interface. On the left, a Digg feed lists several news items, including "Microsoft Demos 'ADD TO DIGG' Feature in IE8", "It was only a matter of time, The SIMS 3 Official", "Universe submerged in a sea of chilled neutrinos", "Unique locks on microchips could reduce hardware piracy", and "Warren Buffett Passes Gates To Become World's Richest Man".

The central section features an artist profile for **Björk**. It includes a search bar for "artist or song name", a "Go" button, and tabs for "Albums" and "Tracks". Below the search bar is a video player showing Björk performing. The profile text identifies her as an Icelandic singer/songwriter and provides her birth date (November 21, 1965) and location (Reykjavík, Iceland). It also includes social sharing options for "Share this artist" and "Blog this artist", and a list of genres: Pop, Trip-Hop, Rock, Vocal Jazz, Ambient, Electronica, Dance, Alternative, and Experimental.

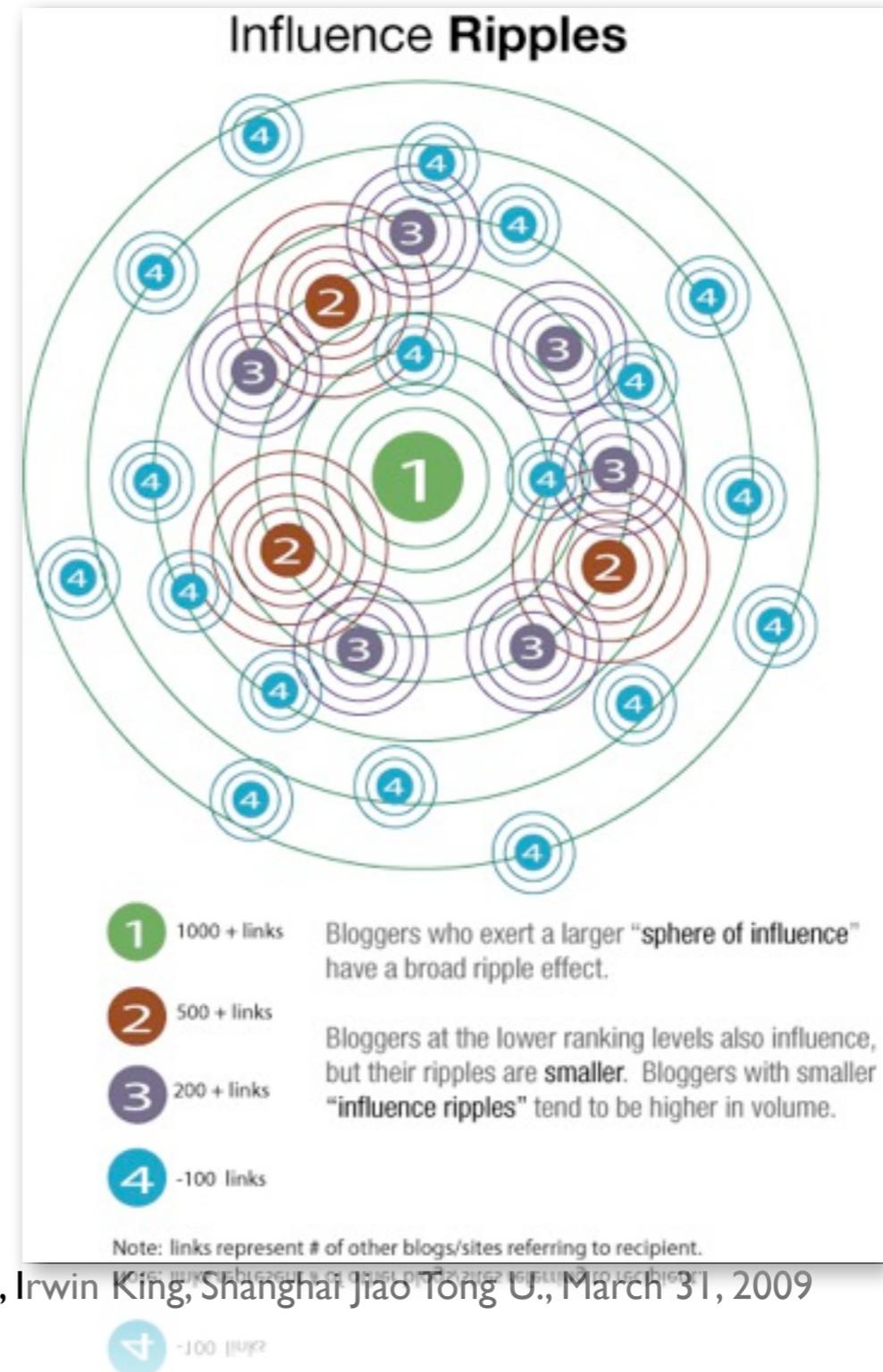
On the right side, there are several music-related widgets:

- Videos on YouTube:** Lists videos such as "All is full of love" (4:09), "bjork-hunter" (3:38), and "Bjork - Human Behaviour" (4:17).
- Lyrics from Yahoo! Music:** Provides a list of songs by track and album, including "5 Years", "Alarm Call", "All Is Full of Love", "All Neon Like", "An Echo, A Stain", "Army of Me", "Aurora", "Bachelorette", "Big Time Sensuality", "Cetacea", "Cocoon", "Come to Me", "Crying", and "Desired Constellation".
- Flickr Photos:** Displays a selection of photos from Flickr, with a "next" button.
- Music on Hype Machine:** Includes a "Play All" button.
- Artist on Last.fm:** Shows a profile for "The Sugarcubes" with a "Goldfrapp" recommendation.



Social Marketing

- Viral marketing
- Who are the **brokers**?
- Who can exert the **most influence** on buying/selling?
- How **much** should one advertise?



Social/Human Computation

Security Check: Enter **both** words below, separated by a space. What's This?
Can't read this? Try another.
[Try an audio captcha](#)

discharge **carolina**

Text in the box:

I have read and agree to the [Terms of Use and Privacy Policy](#)

[Sign Up](#)

[Problems signing up? Check out our help pages](#)

Security Check: Enter **both** words below, separated by a space. What's This?
Can't read this? Try another.
[Try an audio captcha](#)

discharge **tesbiten**

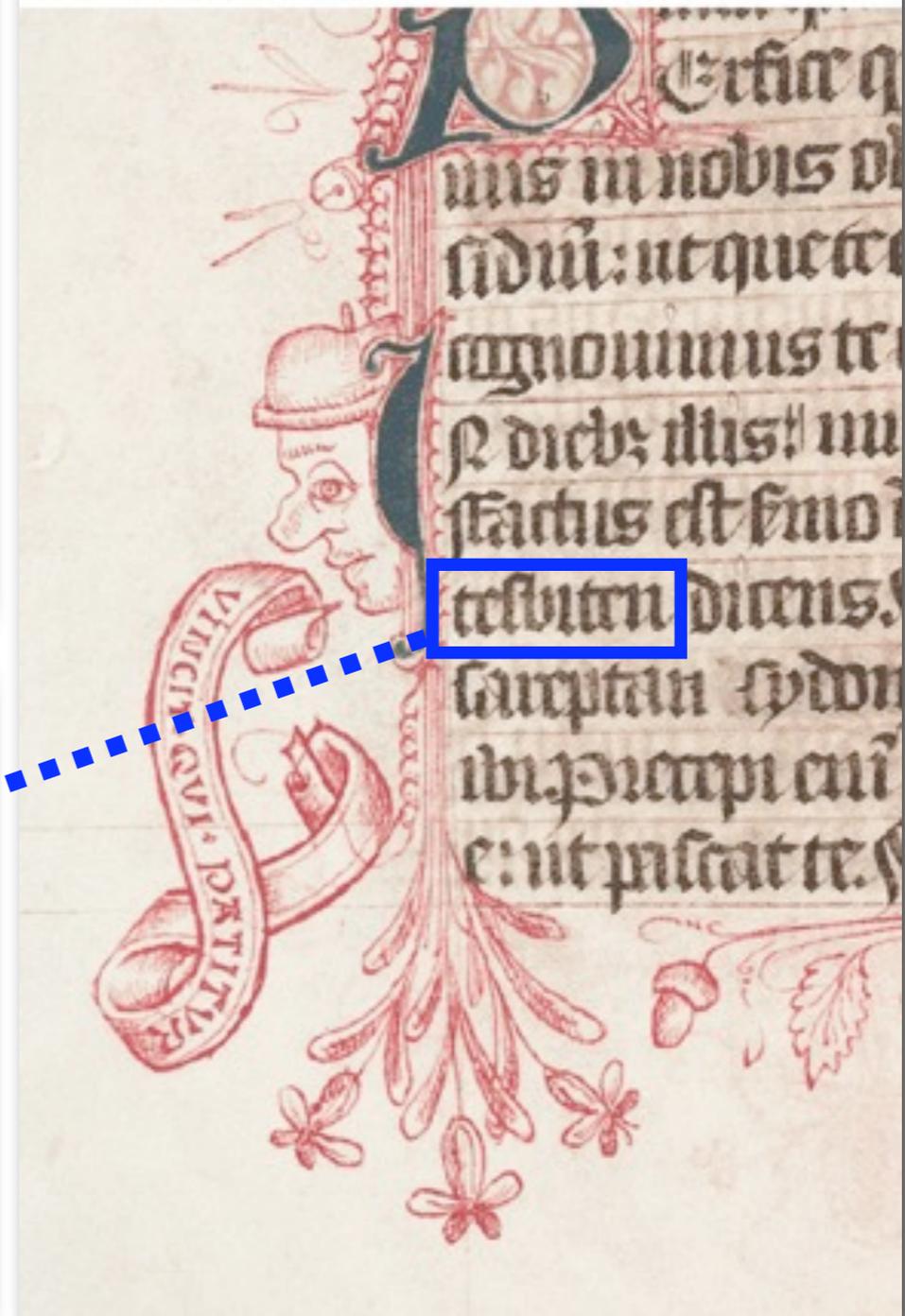
Text in the box:

I have read and agree to the [Terms of Use and Privacy Policy](#)

[Sign Up](#)

[Problems signing up? Check out our help pages](#)

MS. Don. b. 6, fol. 48v (detail) © Bodleian Library, University of Oxford



Human Computation

The screenshot shows the Google Image Labeler interface. At the top left is the Google logo and 'Image Labeler BETA'. The main title is 'Google Image Labeler'. On the left side, there is a scorecard with 'time left' at 01:17, 'score' at 0, and 'passes' at 0. In the center, there is a text box with 'Your partner has suggested 10 labels.' and a 'label' button. Below this is a photograph of a lake and mountains. At the bottom of the photo is a 'zoom out' button. On the right side, there is a list of labels: 'off-limits', 'sky', 'water', 'blue', 'lake', 'mountain', and 'my labels'. At the bottom of the interface are links for 'Privacy Policy', 'Terms of Use', and 'Return to Google Image Search', along with the copyright notice '© 2007 Google'. Several red starburst shapes are overlaid on the interface, highlighting the scorecard, the partner's suggestion, the image, and the list of labels.

© 2007 Google



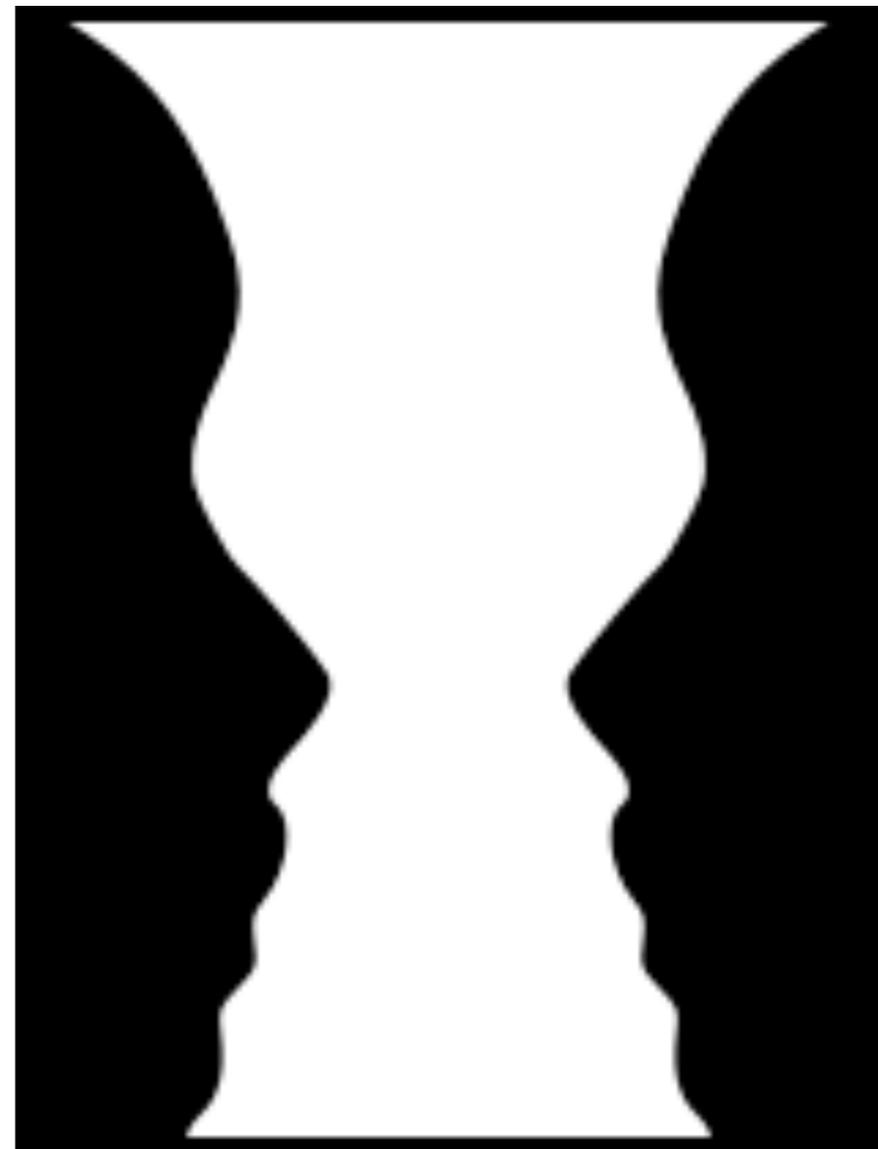
Web 2.0 Revolution

The Three C's

Connectivity

Collaboration

Communities



What's On the Menu?

- Web 2.0 and Social X
- **Social Computing**
- Some Interesting Problems
 - Collaborative Filtering
 - Query Suggestion

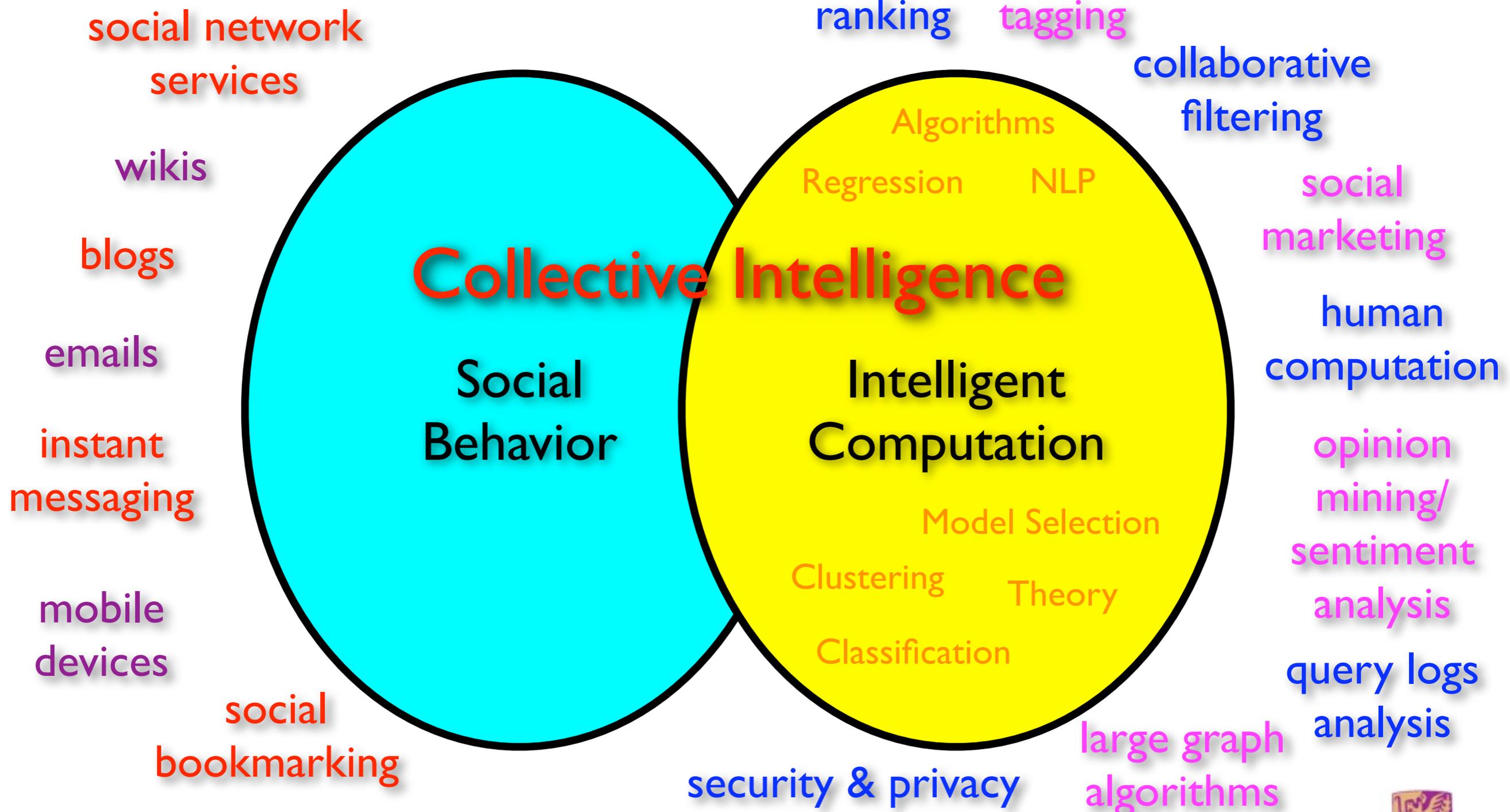


Social Relations

presence
identity
social role
reputation
expertise
trust
ownership
accountability
knowledge
crew
teams
populations
binary
cardinal
integer
real
squad
organizations
cohorts
markets
communities
partners
groups



Social Computing



Social Computing (SC)

- Social computing is a general term for an area of computer science that is concerned with the intersection of *social behavior* and *computational systems*.
Wikipedia
- *A social structure in which technology puts power in communities, not institutions.*
Forrester
- *Forms of web services where the value is created by the collective contributions of a user population.*



Issues

- **Theory** and models
- **Search, mining, and ranking** of existing information, e.g., spatial (relations) and temporal (time) domains
- Dealing with **partial** and **incomplete** information, e.g., collaborative filtering, ranking, tagging, etc.
- **Scalability** and algorithmic issues
- **Security** and **privacy** issues
- **Monetization** of social interactions

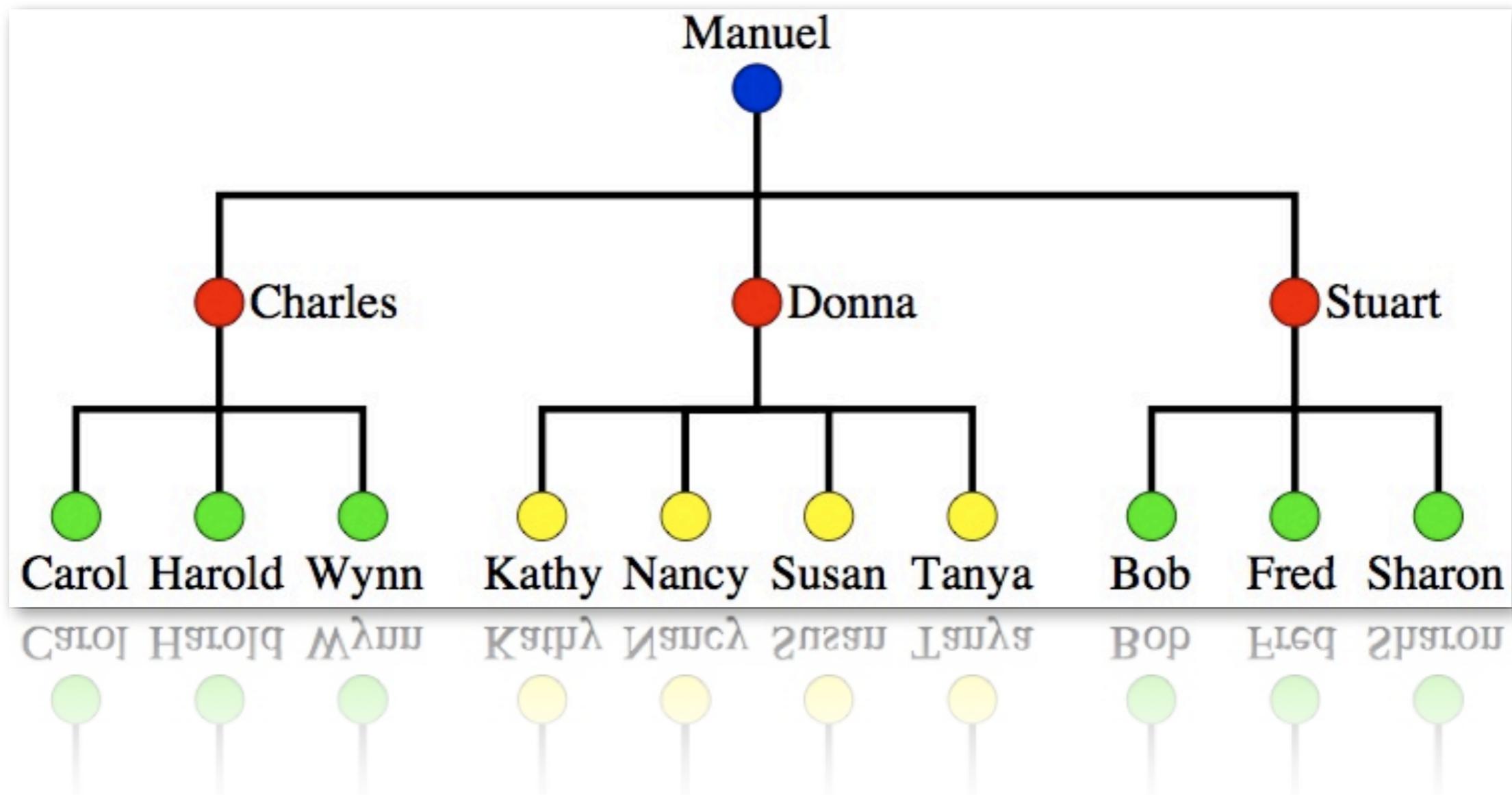


Machine Learning in SC

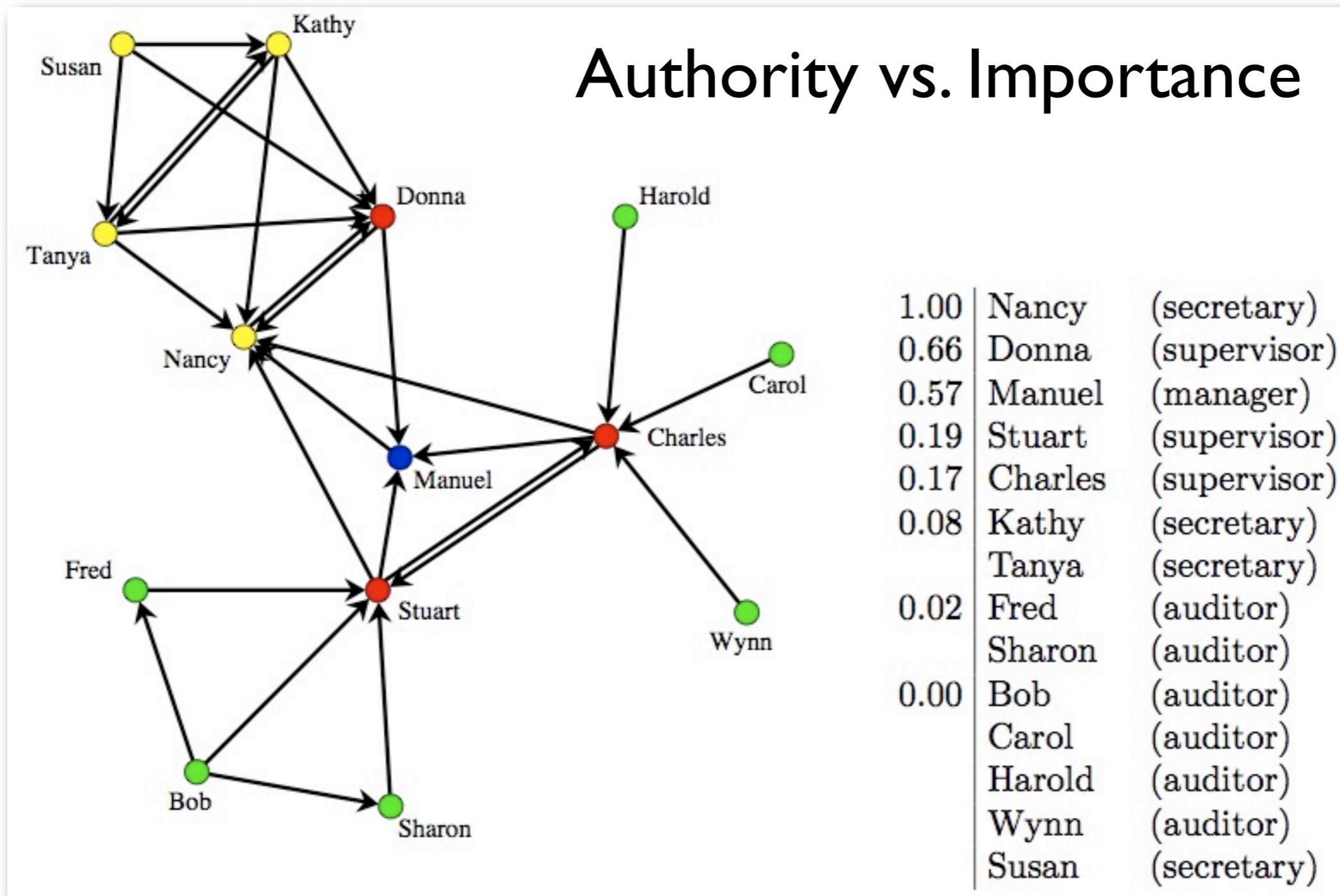
- Classification, clustering, regression, etc.
- New insights on the data
 - Social relations are often **hidden** (latent)
 - Change data from (x, y) to $(x, c_1(x), c_2(x), \dots, y)$
- $c(x)$ = context in **tags, relations, ratings**, etc.
- data type = *binary, integer, real, cardinal*, etc.



Organizational Chart



Social Network Chart



What's On the Menu?

- Web 2.0 and Social X
- Social Computing
- Some Interesting Problems
 - Collaborative Filtering
 - Query Suggestion



What's On the Menu?

- Web 2.0 and Social X
- Social Computing
- Some Interesting Problems
 - Collaborative Filtering
 - Query Suggestion



A Better Mousetrap?

cuil

how to be

- How to Be a Player
- How to be Happy
- How to Be Popular
- How to Be a DJ
- How to Be Cool
- How to be Rich

617,892,992 web pages

Out Cuil | Your Privacy

© 2008 Cuil, Inc

Yahoo! Mail

Web Images Video Local Shopping more

dudley man

YAHOO!

dudley manlove

dudley manlove quartet

dudley mansion

dudley management

dudley mandy

- Muppets**
www.Target.com - Find muppets Online. Shop and Save at Target Today.
- Muppet Show Ringtones**
www.hiptunez.com - Download Muppet Show ringtones to your phone today.

1. **Muppets.com**
Official site of Jim Henson's Muppets, with games, music, downloads, news, and information about all the characters.
muppets.go.com - 4k - Cached

Google

Live Search | MSN | Windows Live | Hotmail

United States | Options | cashback | Sign in

Live Search apple

Web 1-10 of 132,000,000 results - Advanced
See also: Images, Video, News, Maps, More

Apple Items - www.ebay.com Sponsored sites
Buy Apple Items. You may get 25% off with PayPal if eligible.

macintosh - Search.Live.com/cashback
Earn cashback on millions of products from sites you trust!

Official Apple Store - store.apple.com
Buy the new MacBook, Air, and Pro direct from Apple. Free shipping.

Apple
Apple designs and creates iPod and iTunes, Mac laptop and desktop computers, the OS X operating system, and the revolutionary iPhone.
www.apple.com - Cached page

iPod+iTunes
The Apple Store
Downloads
Support
Show more results from www.apple.com

Related searches

Apple iPod
iTunes
Apple Computers
Apple Vacations
Apple Store
Apple Trailers
Best Buy

Sponsored sites
You're a PC Too
Post a Picture. Tell Us How You Use Your PC. See Yourself in the Ads.
ImAPC.LifeWithoutWalls.com

See your message here...

news, fan forum, and Muppet Show episode guide.

al.com liberal a

liberal arts	13,100,000 results
liberal arts colleges	818,000 results
liberal arts college	2,460,000 results
liberal arts degree	613,000 results
liberal arts education	1,710,000 results
liberal arts schools	1,320,000 results
liberal arts major	6,130,000 results
liberal arts college rankings	341,000 results
liberal art	2,950,000 results
liberal arts college ranking	457,000 results

[close](#)



Challenges

- Queries contain **ambiguous** and **new** terms
- **apple**: “apple computer” or “apple pie”?
- **NDCG**:?
- Users tend to submit **short queries** consisting of only one or two words
- almost **20%** one-word queries
- almost **30%** two-word queries

- Users may have **little or even no knowledge** about the topic they are searching for!



What is Clickthrough Data

- Query logs recorded by search engines

$$\langle u, q, l, r, t \rangle$$

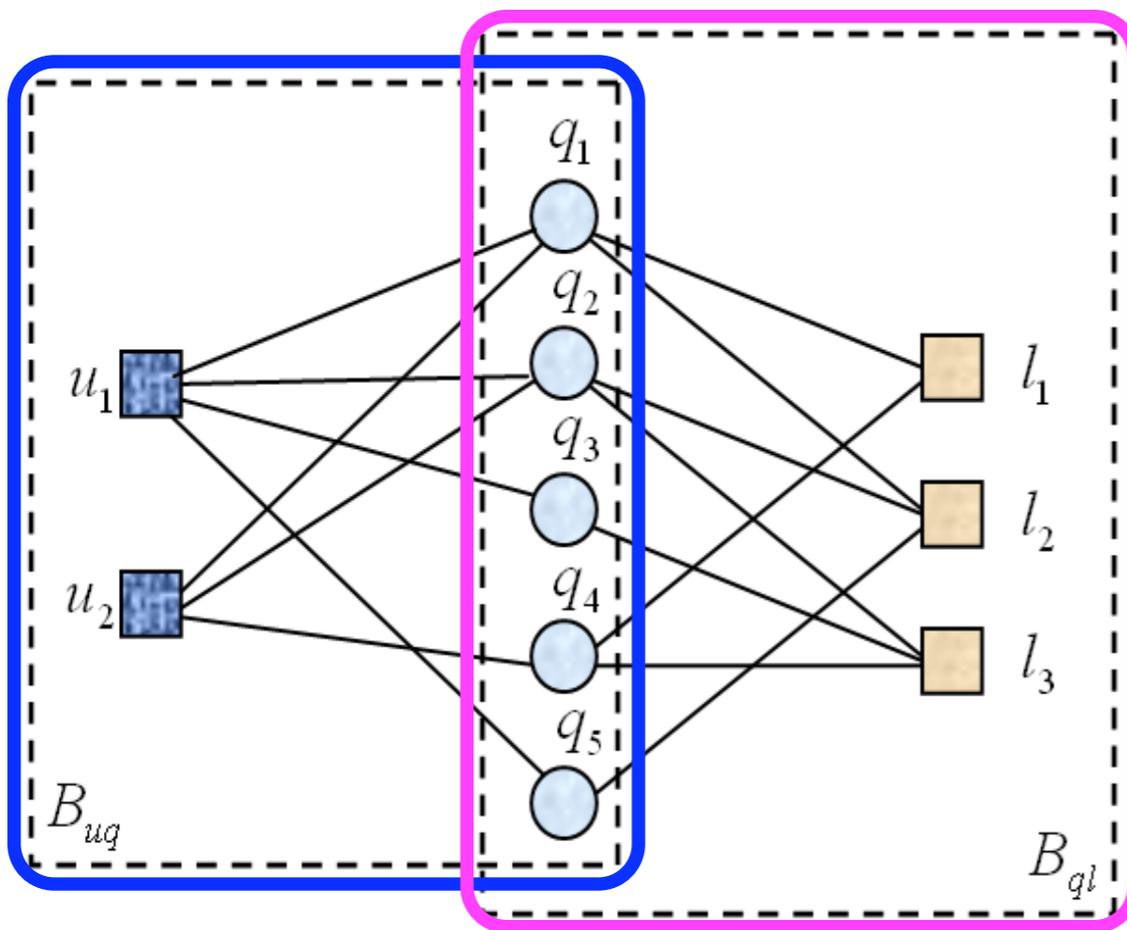
Table 1: Samples of search engine clickthrough data

ID	Query	URL	Rank	Time
358	facebook	http://www.facebook.com	1	2008-01-01 07:17:12
358	facebook	http://en.wikipedia.org/wiki/Facebook	3	2008-01-01 07:19:18
3968	apple iphone	http://www.apple.com/iphone/	1	2008-01-01 07:20:36
...

- Users' **relevance feedback** to indicate desired/preferred/target results



Joint Bipartite Graph



$$B_{uq} = (V_{uq}, E_{uq})$$

$$V_{uq} = U \cup Q$$

$$U = \{u_1, u_2, \dots, u_m\}$$

$$Q = \{q_1, q_2, \dots, q_n\}$$

$E_{uq} = \{(u_i, q_j) \mid \text{there is an edge from } u_i \text{ to } q_j\}$
is the set of all edges.

The edge (u_i, q_j) exists in this bipartite graph if and only if a user u_i issued a query q_j .

$$B_{ql} = (V_{ql}, E_{ql})$$

$$V_{ql} = Q \cup L$$

$$Q = \{q_1, q_2, \dots, q_n\}$$

$$L = \{l_1, l_2, \dots, l_p\}$$

$E_{ql} = \{(q_i, l_j) \mid \text{there is an edge from } q_i \text{ to } l_j\}$
is the set of all edges.

The edge (q_j, l_k) exists if and only if a user u_i clicked a URL l_k after issuing an query q_j .



Key Points

- Two-level latent semantic analysis

Level
1

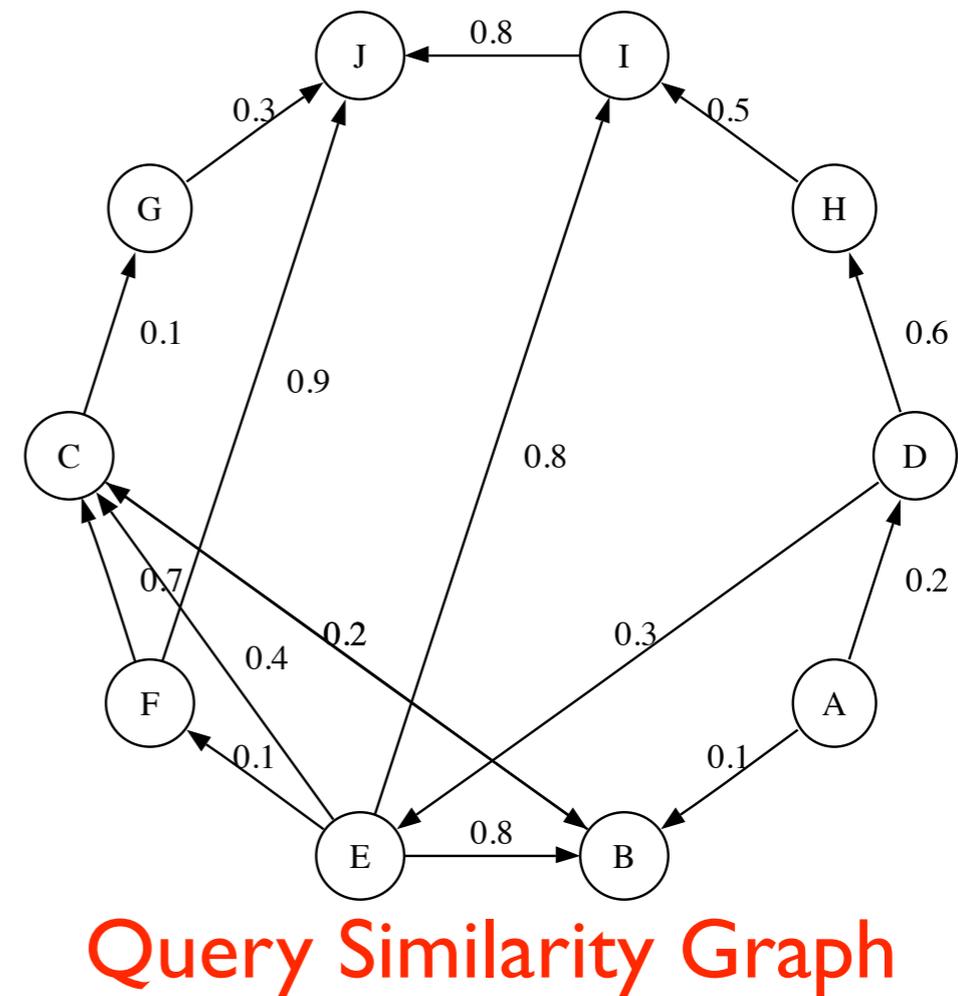
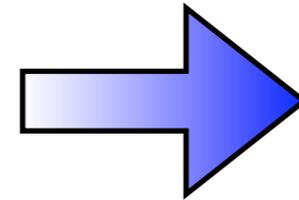
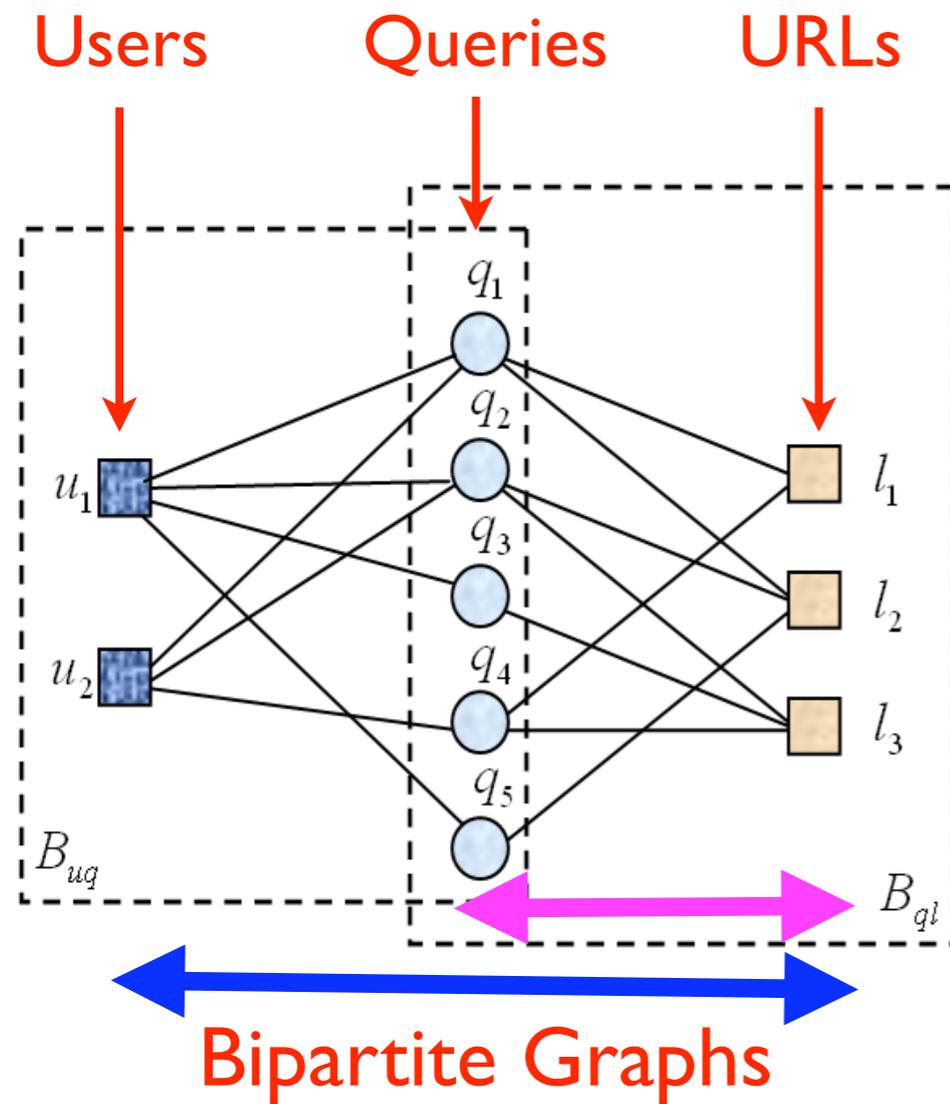
- Consider the use of a joint **user-query** and **query-URL bipartite graphs** for query suggestion

Level
2

- Use **matrix factorization** for learning query features in constructing the Query Similarity Graph

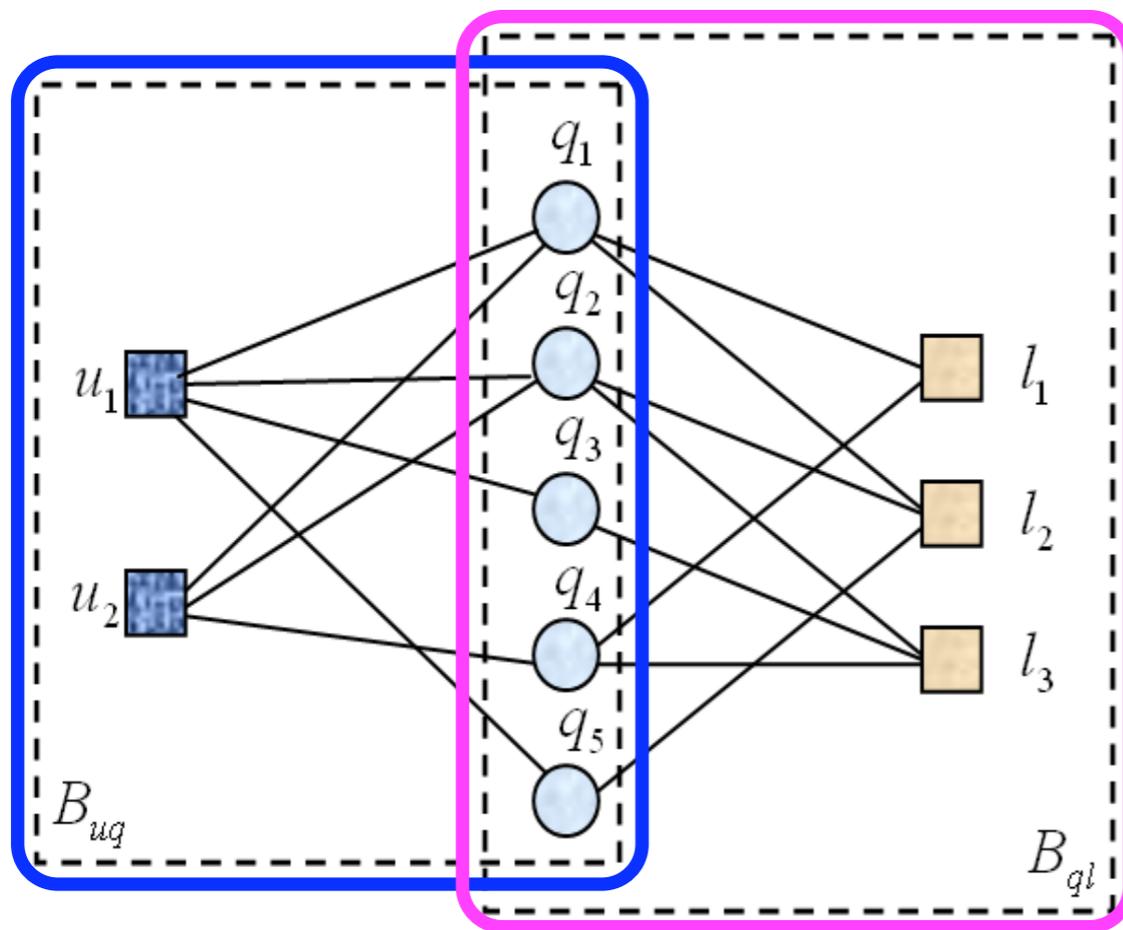
- Use **heat diffusion** for similarity propagation for query suggestions





- Queries are issued by the users, and which URLs to click are also decided by the users
- Two distinct users are similar if they issued **similar queries**
- Two queries are similar if they are issued by **similar users**



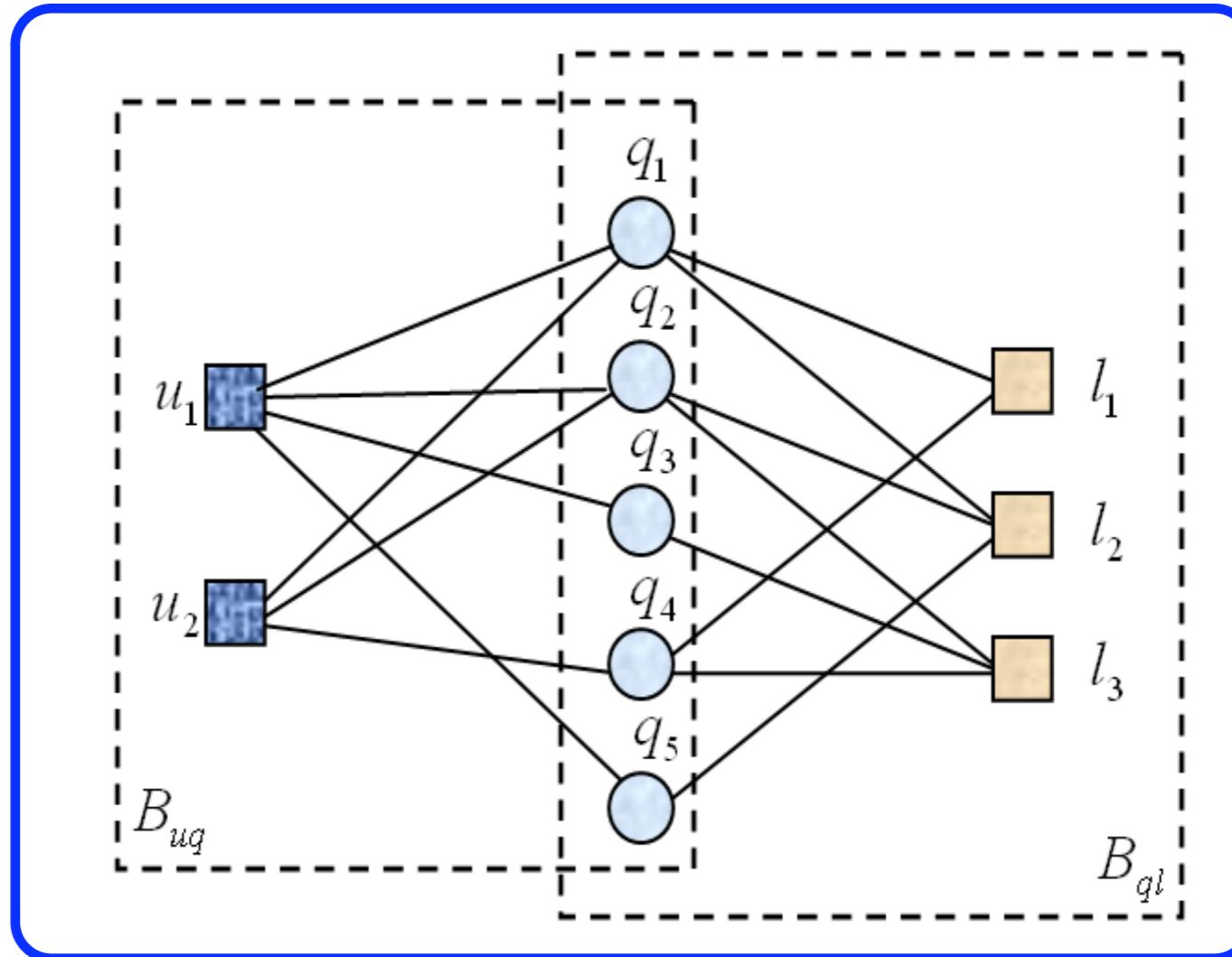


- r_{ij}^* Normalized weight, how many times u_i issued q_j
- s_{jk}^* Normalized weight, how many times q_j is linked to l_k
- U_i L -dimensional vector of user u_i
- Q_j L -dimensional vector of query q_j
- L_k L -dimensional vector of URL l_k

$$\mathcal{H}(R, U, Q) = \min_{U, Q} \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^n I_{ij}^R (r_{ij}^* - g(U_i^T Q_j))^2 + \frac{\alpha_u}{2} \|U\|_F^2 + \frac{\alpha_q}{2} \|Q\|_F^2$$

$$\mathcal{H}(S, Q, L) = \min_{Q, L} \frac{1}{2} \sum_{j=1}^n \sum_{k=1}^p I_{jk}^S (s_{jk}^* - g(Q_j^T L_k))^2 + \frac{\alpha_q}{2} \|Q\|_F^2 + \frac{\alpha_l}{2} \|L\|_F^2$$





$$\mathcal{H}(S, R, U, Q, L) =$$

$$\frac{1}{2} \sum_{j=1}^n \sum_{k=1}^p I_{jk}^S (s_{jk}^* - g(Q_j^T L_k))^2 + \frac{\alpha_r}{2} \sum_{i=1}^m \sum_{j=1}^n I_{ij}^R (r_{ij}^* - g(U_i^T Q_j))^2$$

$$+ \frac{\alpha_u}{2} \|U\|_F^2 + \frac{\alpha_q}{2} \|Q\|_F^2 + \frac{\alpha_l}{2} \|L\|_F^2,$$

- A local minimum can be found by performing **gradient descent** in U_i , Q_j and L_k



Gradient Descent Equations

$$\frac{\partial \mathcal{H}}{\partial U_i} = \alpha_r \sum_{j=1}^n I_{ij}^R g'(U_i^T Q_j) (g(U_i^T Q_j) - r_{ij}^*) Q_j + \alpha_u U_i,$$

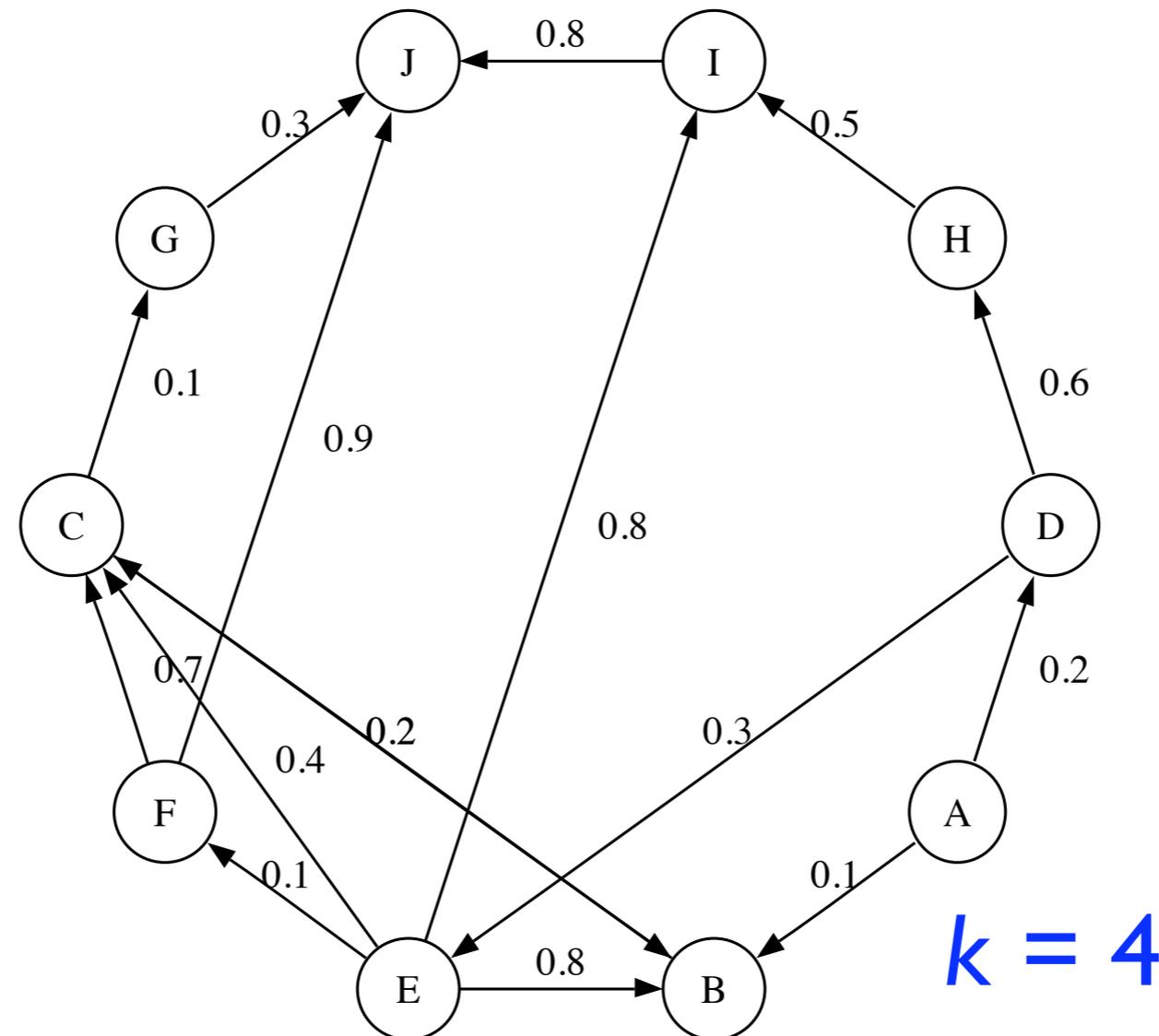
$$\begin{aligned} \frac{\partial \mathcal{H}}{\partial Q_j} &= \sum_{k=1}^p I_{jk}^S g'(Q_j^T L_k) (g(Q_j^T L_k) - s_{jk}^*) L_k \\ &+ \alpha_r \sum_{i=1}^m I_{ij}^R g'(U_i^T Q_j) (g(U_i^T Q_j) - r_{ij}^*) U_i + \alpha_q Q_j, \end{aligned}$$

$$\frac{\partial \mathcal{H}}{\partial L_k} = \sum_{j=1}^n I_{jk}^S g'(Q_j^T L_k) (g(Q_j^T L_k) - s_{jk}^*) Q_j + \alpha_l L_k,$$

Only the **Q matrix**, the queries' latent features, is being used to generate the **query similarity graph!**



Query Similarity Graph



- Similarities are calculated using queries' latent features
- Only the **top- k** similar neighbors (terms) are kept



Similarity Propagation

- Based on the **Heat Diffusion Model**
- In the query graph, given the **heat sources** and the **initial heat values**, start the heat diffusion process and perform **P steps**
- Return the **Top- N** queries in terms of highest heat values for query suggestions



Heat Diffusion Model

- Heat diffusion is a **physical phenomena**
- Heat flows from **high** temperature to **low** temperature in a **medium**
- **Heat kernel** is used to describe the amount of heat that one point receives from another point
- The way that heat diffuse varies when the **underlying geometry**

$$\rho C_P \frac{\partial T}{\partial t} = Q + \nabla \cdot (k \nabla T)$$

ρ Density

C_P Heat capacity and constant pressure

$\frac{\partial T}{\partial t}$ Change in temperature over time

Q Heat added

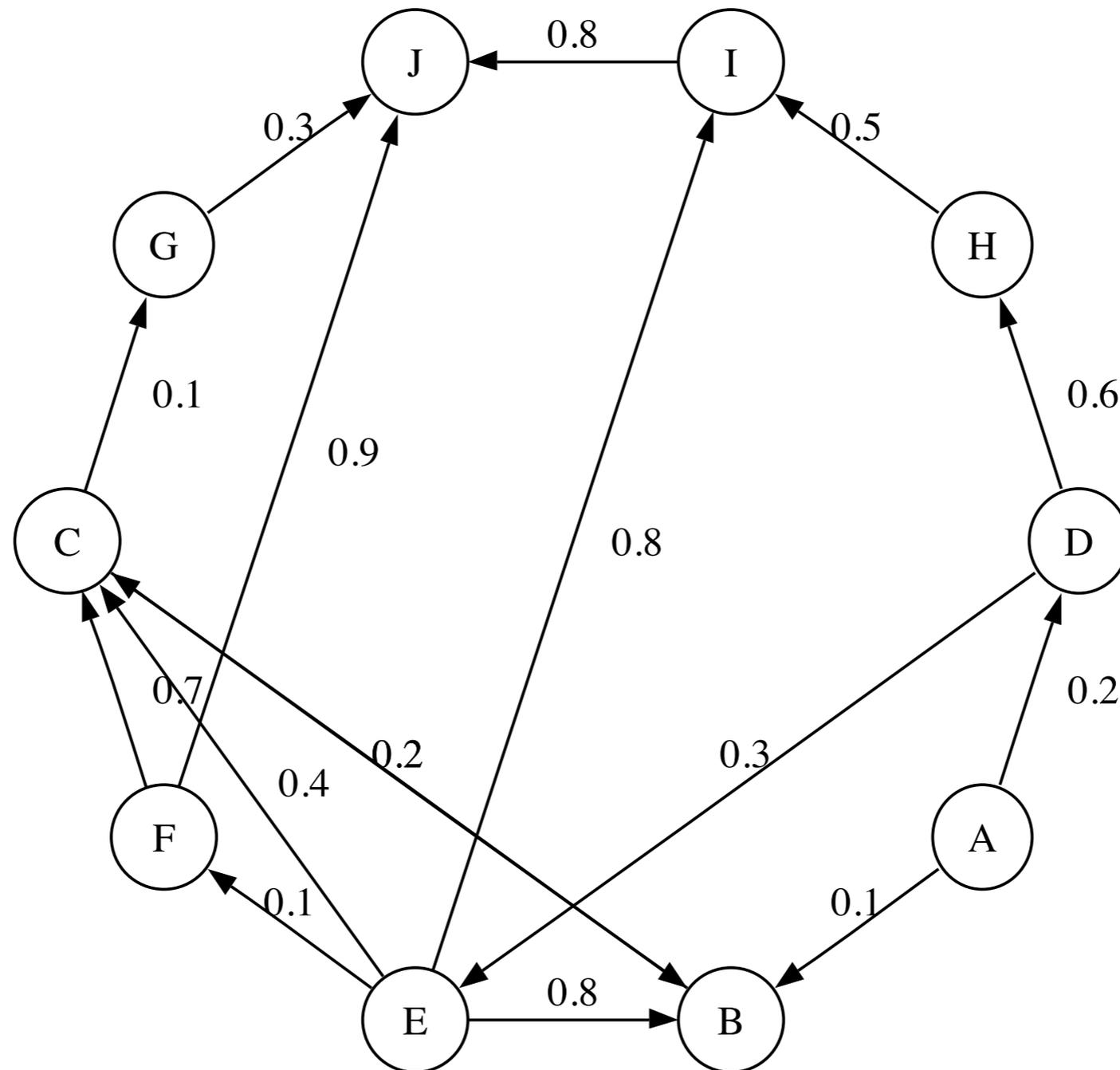
k Thermal conductivity

∇T Temperature gradient

$\nabla \cdot \mathbf{v}$ Divergence



Heat Diffusion Process



Similarity Propagation Model

$$\frac{f_i(t + \Delta t) - f_i(t)}{\Delta t} = \alpha \left(-\frac{\tau_i}{d_i} f_i(t) \sum_{k:(q_i, q_k) \in E} w_{ik} + \sum_{j:(q_j, q_i) \in E} \frac{w_{ji}}{d_j} f_j(t) \right) \quad (1)$$

$$\mathbf{f}(1) = e^{\alpha \mathbf{H}} \mathbf{f}(0) \quad (2)$$

$$H_{ij} = \begin{cases} w_{ji}/d_j, & (q_j, q_i) \in E, \\ -(\tau_i/d_i) \sum_{k:(i,k) \in E} w_{ik}, & i = j, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

$$\mathbf{f}(1) = e^{\alpha \mathbf{R}} \mathbf{f}(0), \quad \mathbf{R} = \gamma \mathbf{H} + (1 - \gamma) \mathbf{g} \mathbf{1}^T \quad (4)$$

- α Thermal conductivity
- d_i Heat value of node i at time t
- $f_i(t)$ Heat value of node i at time t
- w_{ik} Weight between node i and node k
- $\mathbf{f}(0)$ Vector of the initial heat distribution
- $\mathbf{f}(1)$ Vector of the heat distribution at time 1
- τ_i Equal to 1 if node i has outlinks, else equal to 0
- γ Random jump parameter, and set to 0.85
- \mathbf{g} Uniform stochastic distribution vector



Discrete Approximation

- Compute $e^{\alpha \mathbf{R}}$ is time consuming
- We use the **discrete approximation** to substitute

$$\mathbf{f}(1) = \left(\mathbf{I} + \frac{\alpha}{P} \mathbf{R} \right)^P \mathbf{f}(0)$$

- For every heat source, only diffuse heat to its neighbors within **P steps**
- In our experiments, $P = 3$ already generates fairly good results



Query Suggestion Procedure

- For a given query q
 1. Select a set of n queries, each of which contains at least one word in common with q , as **heat sources**

2. Calculate the initial heat values by

$$f_{\hat{q}_i}(0) = \frac{|\mathcal{W}(q) \cap \mathcal{W}(\hat{q}_i)|}{|\mathcal{W}(q) \cup \mathcal{W}(\hat{q}_i)|}$$

$q = \text{"Sony"}$
 $\text{"Sony"} = 1$

$\text{"Sony Electronics"} = 1/2$

$\text{"Sony Vaio Laptop"} = 1/3$

3. Use $\mathbf{f}(1) = e^{\alpha \mathbf{R}} \mathbf{f}(0)$ to diffuse the heat in graph

4. Obtain the **Top-N** queries from $\mathbf{f}(1)$



Physical Meaning of α

- If set α to a large value
 - The results depend more on the query graph, and **more semantically** related to original queries, e.g., **travel => lowest air fare**
- If set α to a small value
 - The results depend more on the initial heat distributions, and **more literally** similar to original queries, e.g., **travel => travel insurance**



Experimental Dataset

Data Source	Clickthrough data from AOL search	After Pre-Processing
Collection Period	March 2006 to May 2006 (3 months)	
Lines of Logs	19,442,629	
Unique user IDs	657,426	192,371
Unique queries	4,802,520	224,165
Unique URLs	1,606,326	343,302
Unique words		69,937



Query Suggestions

Table 2: Examples of LSQS Query Suggestion Results ($k = 50$)

Testing Queries	Suggestions				
	$\alpha = 10$			$\alpha = 1000$	
	Top 1	Top 2	Top 3	Top 4	Top 5
michael jordan	michael jordan shoes	michael jordan bio	pictures of michael jordan	nba playoff	nba standings
travel	travel insurance	abc travel	travel companions	hotel tickets	lowest air fare
java	sun java	java script	java search	sun microsystems inc	virtual machine
global services	ibm global services	global technical services	staffing services	temporary agency	manpower professional
walt disney land	world of disney	disney world orlando	disney world theme park	disneyland grand hotel	disneyland in california
intel	intel vs amd	amd vs intel	pentium d	pentium	centrino
job hunt	jobs in maryland	monster job	jobs in mississippi	work from home online	monster board
photography	photography classes	portrait photography	wedding photography	adobe elements	canon lens
internet explorer	ms internet explorer	internet explorer repair	internet explorer upgrade	microsoft com	security update
fitness	fitness magazine	lifestyles family fitness	fitness connection	womens health magazine	family fitness
m schumacher	schumacher	red bull racing	formula one racing	ferrari cars	formula one
solar system	solar system project	solar system facts	solar system planets	planet jupiter	mars facts
sunglasses	replica sunglasses	cheap sunglasses	discount sunglasses	safilo	marhon
search engine	audio search engine	best search engine	search engine optimization	song lyrics search	search by google
disease	grovers disease	liver disease	morgellons disease	colic in babies	oklahoma vital records
pizzahut	pizza hut menu	pizza coupons	pizza hut coupons	papa johns pizza coupon	papa johns
health care	health care proxy	universal health care	free health care	great west healthcare	uhc
flower delivery	global flower delivery	online florist	flowers online	send flowers	virtual flower
wedding	wedding guide	wedding reception ideas	wedding decoration	unity candle	centerpiece ideas
astronomy	astronomy magazine	astronomy pic of the day	star charts	space pictures	comet



Comparisons

Table 3: Comparisons between LSQS and SimRank

	Top 1	Top 2	Top 3	Top 4	Top 5
jaguar					
LSQS	jaguar cat	jaguar commercial	jaguar parts	jaguarundi	leopard
SimRank	american black bear	bottlenose dolphin	leopard	margay	jaguarundi
apple					
LSQS	apple computers	apple ipod	apple diet	apple vacations	apple bottom
SimRank	ipod troubleshooting	apple quicktime	apple ipods	apple computers	apple software

Table 4: Accuracy Comparisons

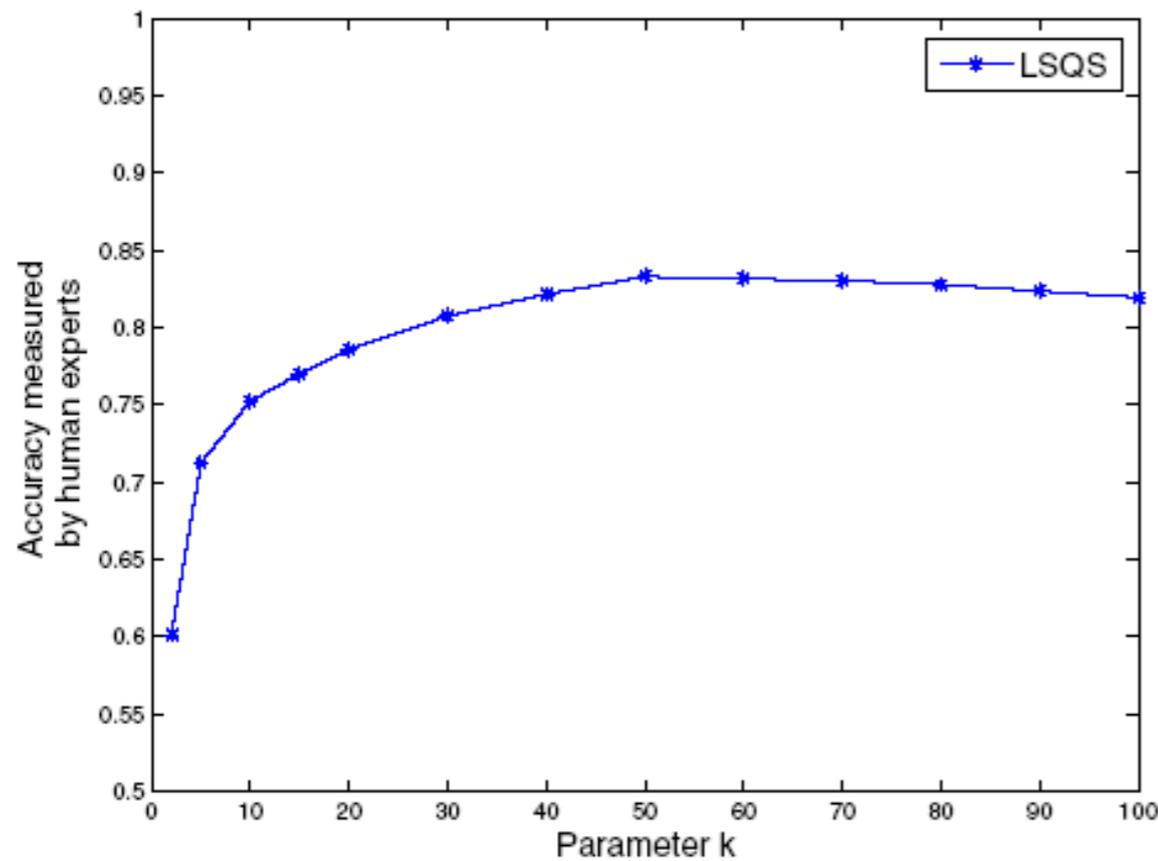
Accuracy	LSQS	SimRank
By Experts	0.8413	0.7101
By ODP	0.6823	0.5789

ODP, Open Directory Project, see <http://dmoz.org>

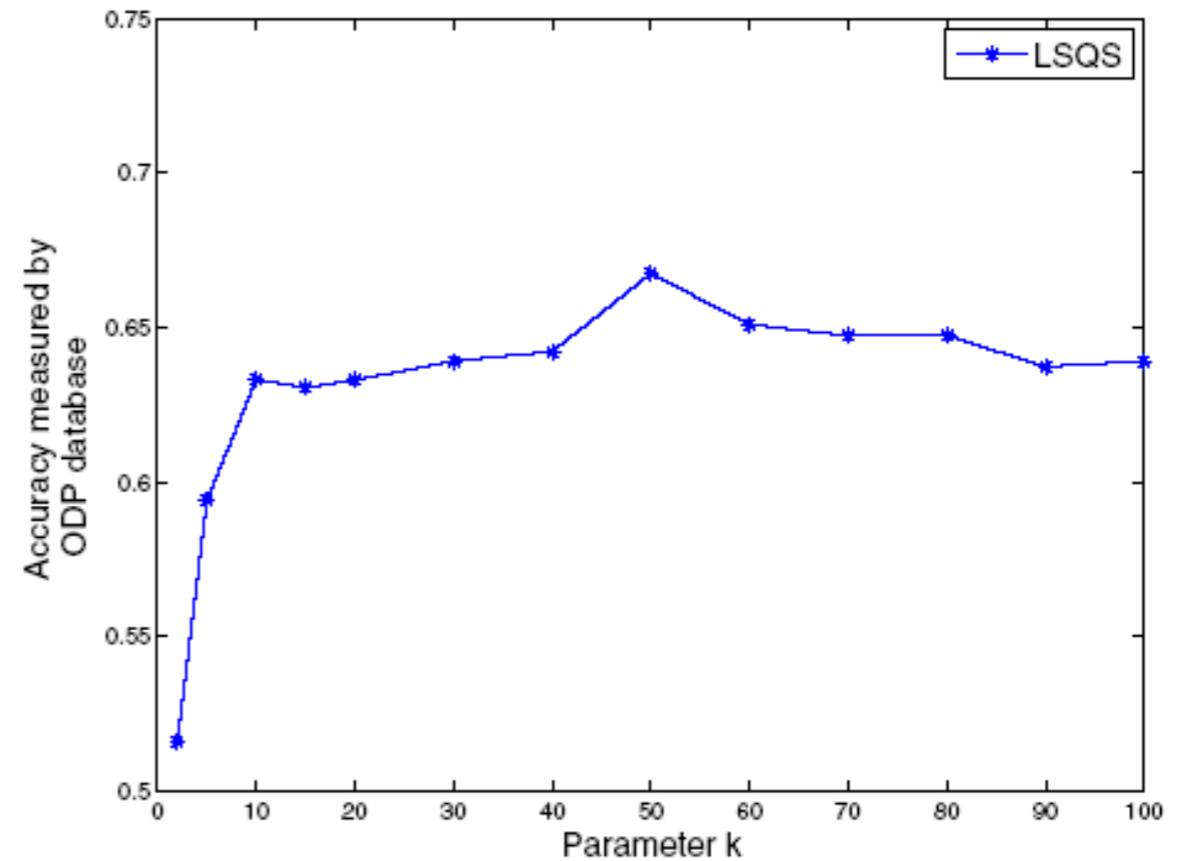


Impact of Parameter k

To test the extend of similarity needed



(a) Evaluation by Experts



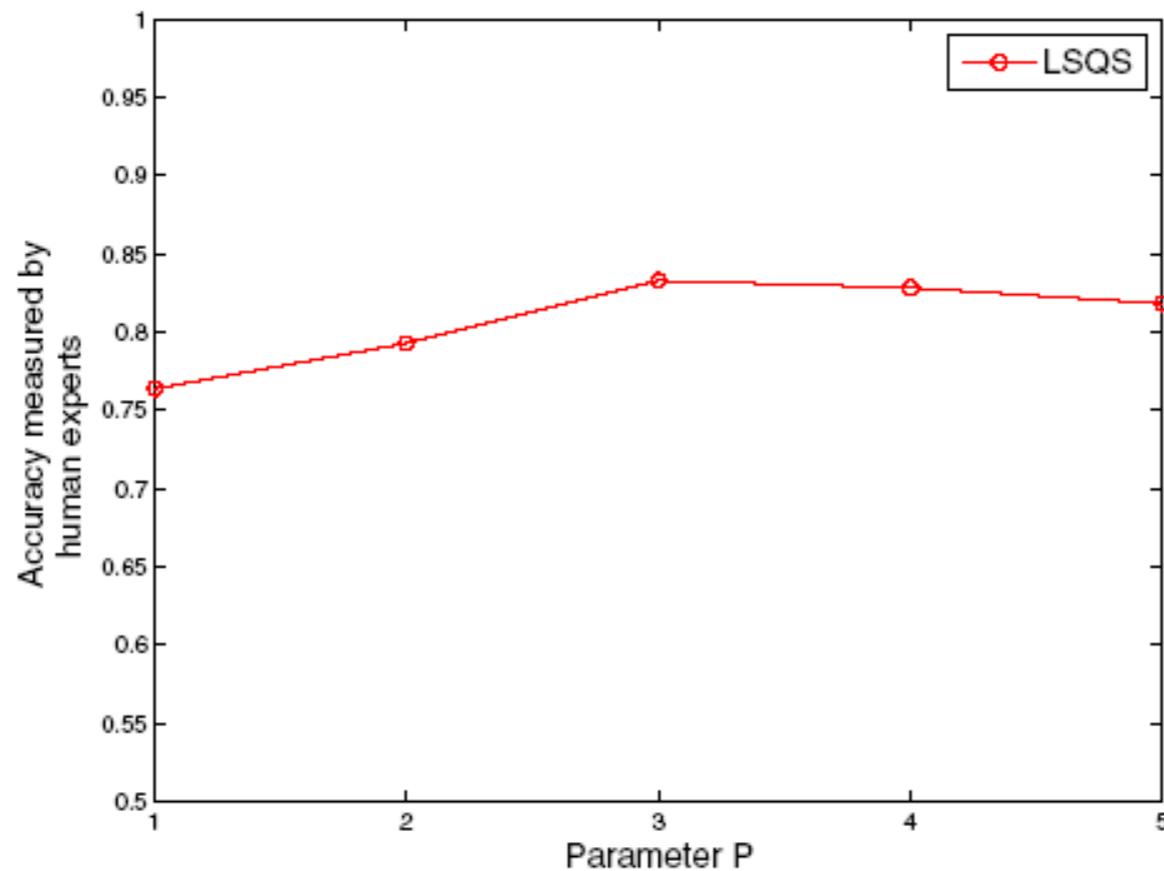
(b) Evaluation by ODP Database

Figure 2: Impact of Parameter k ($P = 3$)

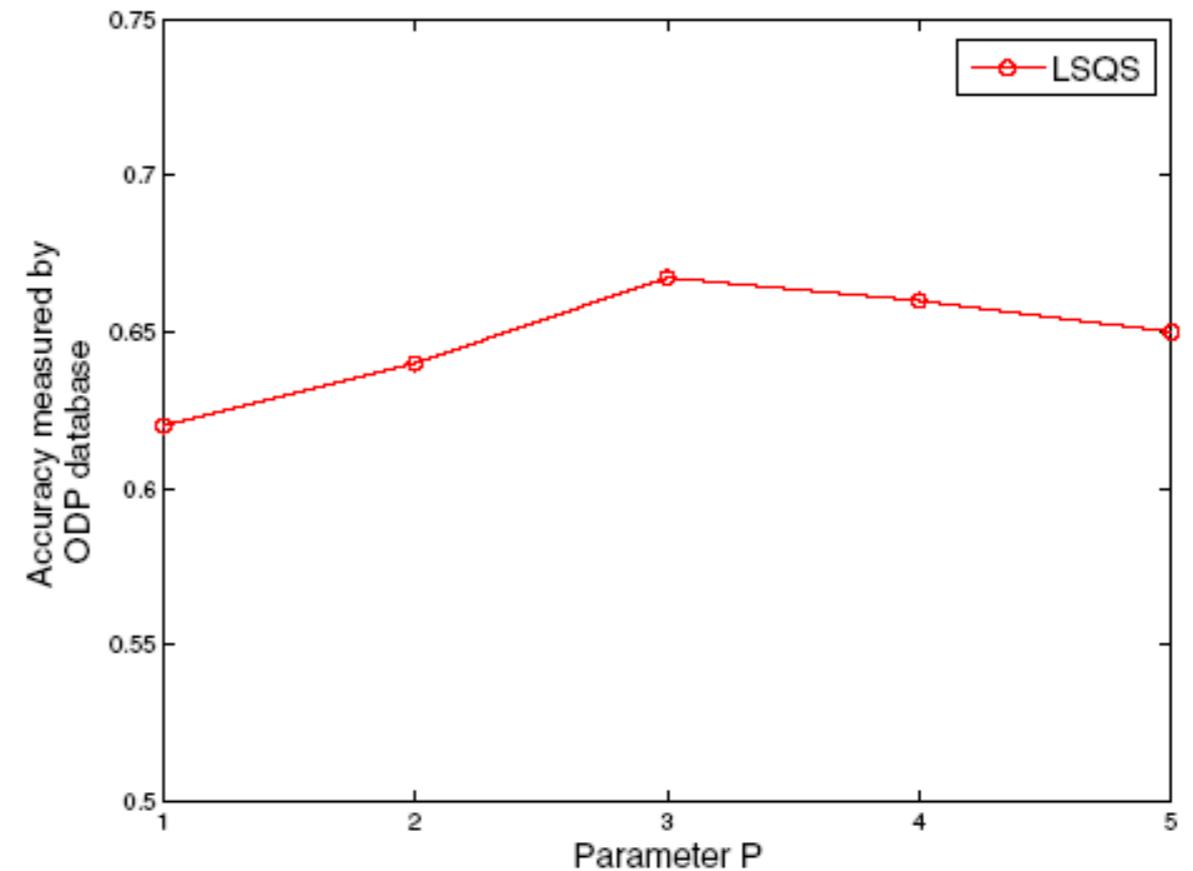


Impact of Parameter P

To test the propagation influence



(a) Evaluation by Experts

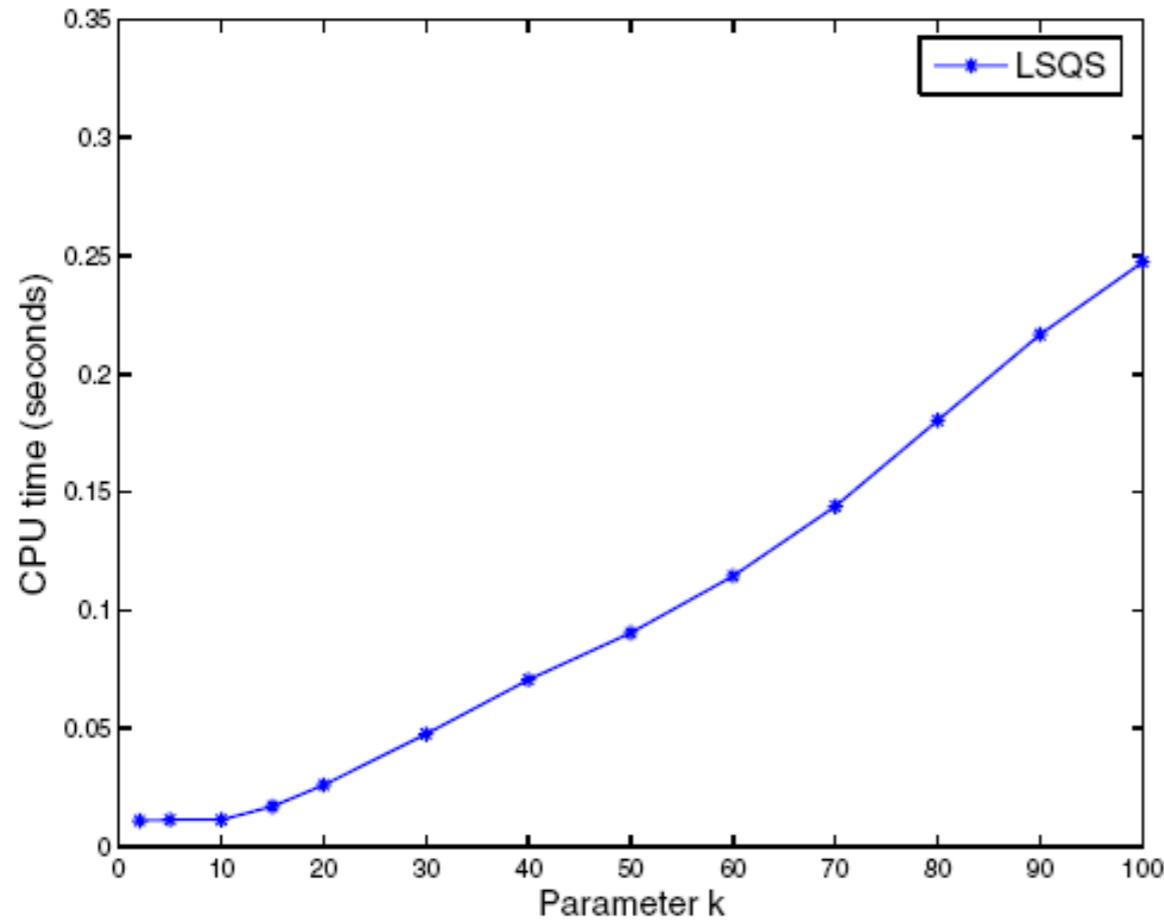


(b) Evaluation by ODP Database

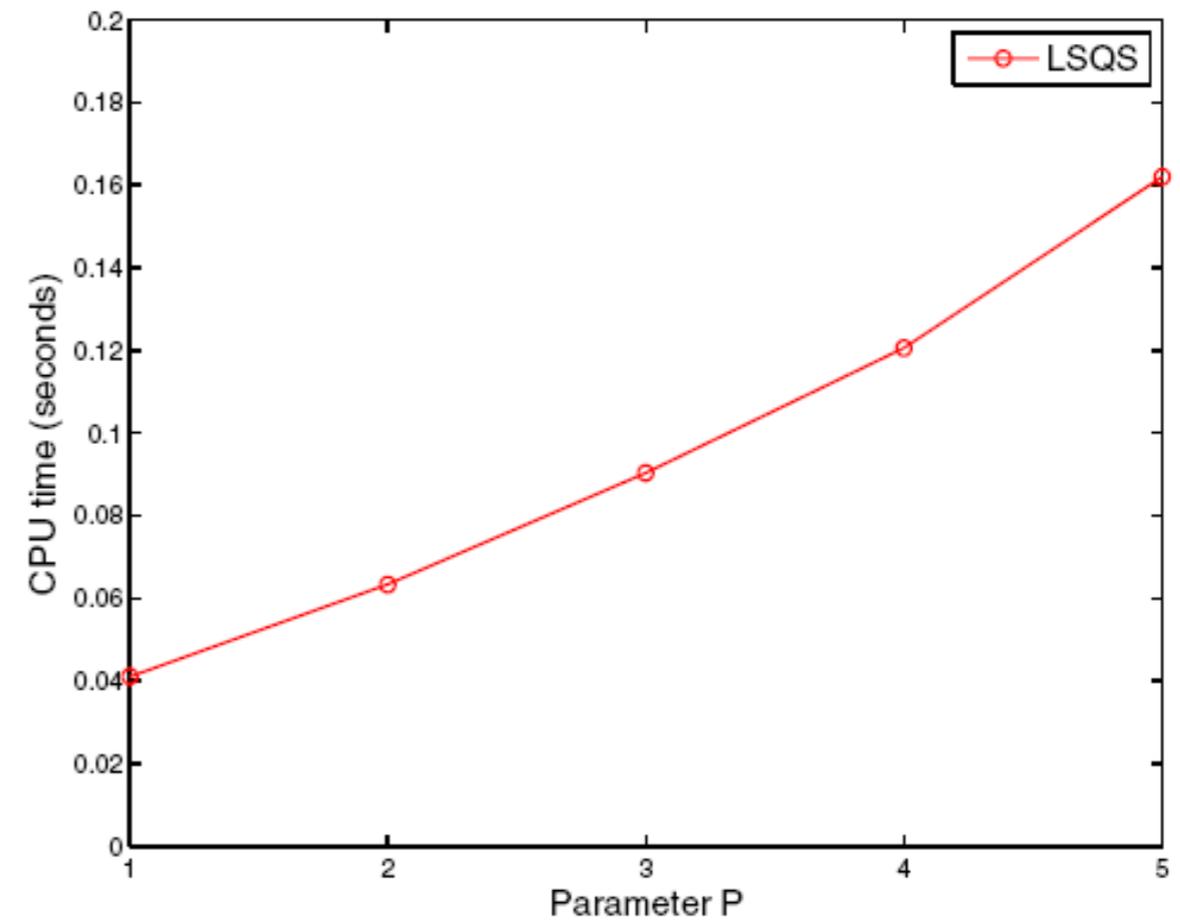
Figure 3: Impact of Parameter P ($k = 50$)



Efficiency Analysis



(a) $P = 3$



(b) $k = 50$

Figure 4: Efficiency Analysis



Summary

- Propose an offline novel **joint matrix factorization** method using **user-query** and **query-URL bipartite graphs** for learning query features
- Propose an online diffusion-based **similarity propagation** and **ranking method** for query suggestion



Conclusion

- Social Computing is a **paradigm shift!**
- Novel views on the **spatial** and **temporal** relationship among **social entities!**
- Great **opportunities** in a new research direction!



On-Going Research

Machine Learning

- Direct Zero-norm Optimization for Feature Selection (ICDM'08)
- Semi-supervised Learning from General Unlabeled Data (ICDM'08)
- Learning with Consistency between Inductive Functions and Kernels (NIPS'08)
- An Extended Level Method for Efficient Multiple Kernel Learning (NIPS'08)
- Semi-supervised Text Categorization by Active Search (CIKM'08)
- Transductive Support Vector Machine (NIPS'07)
- Global and local learning (ICML'04, JMLR'04)

Web Intelligence

- Effective Latent Space Graph-based Re-ranking Model with Global Consistency (WSDM'09)
- Formal Models for Expert Finding on DBLP Bibliography Data (ICDM'08)

- Learning Latent Semantic Relations from Query Logs for Query Suggestion (CIKM'08)
- RATE: a Review of Reviewers in a Manuscript Review Process (WI'08)
- MatchSim: link-based web page similarity measurements (WI'07)
- Diffusion rank: Ranking web pages based on heat diffusion equations (SIGIR'07)
- Web text classification (WWW'07)

Collaborative Filtering

- Recommender system: accurate recommendation based on sparse matrix (SIGIR'07)
- SoRec: Social Recommendation Using Probabilistic Matrix Factorization (CIKM'08)

Human Computation

- An Analytical Study of Puzzle Selection Strategies for the ESP Game (WI'08)
- An Analytical Approach to Optimizing The Utility of ESP Games (WI'08)



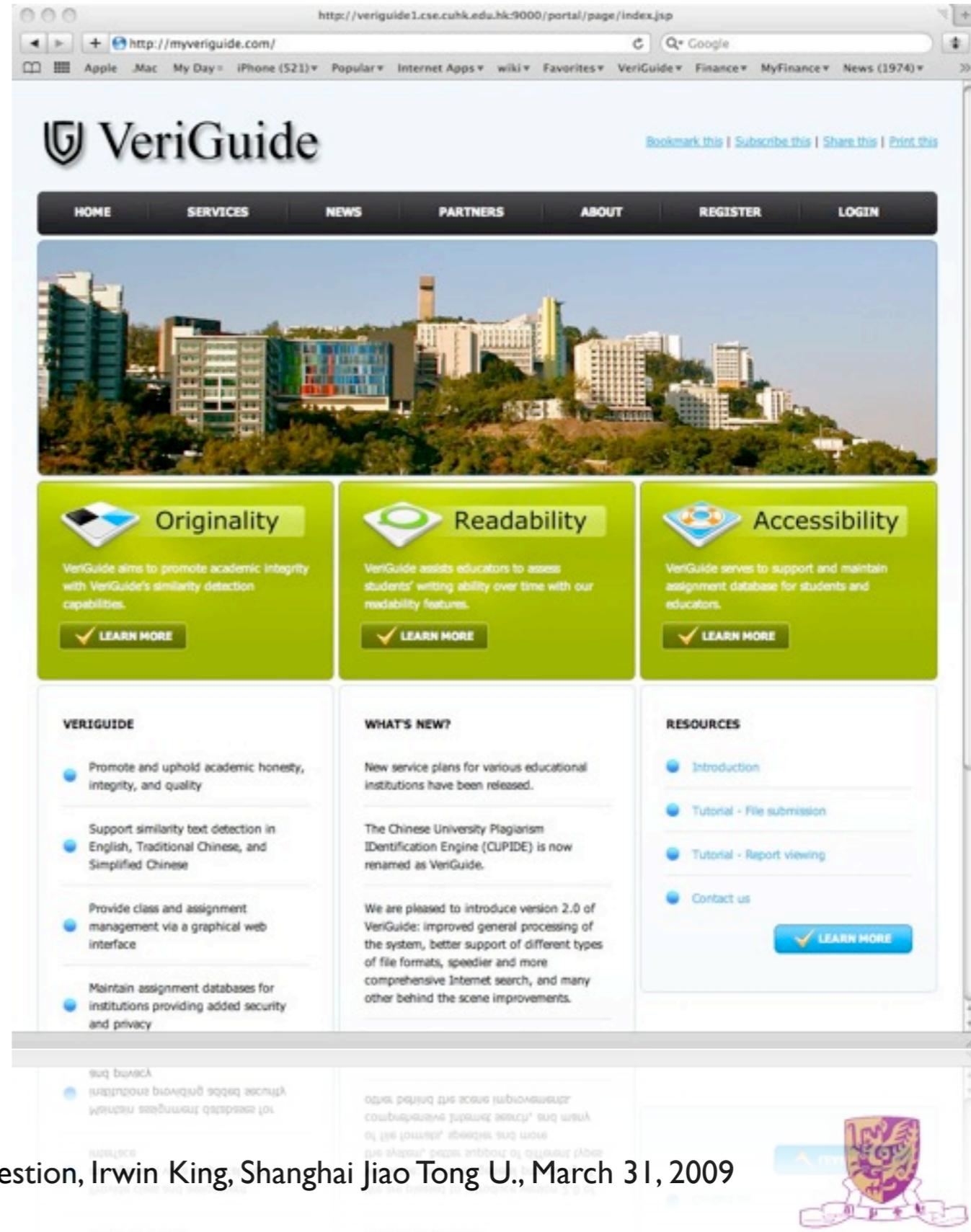
Acknowledgments

- Prof. Michael R. Lyu
- Prof. Jimmy Lee
- Dr. Kaizhu Huang
- Dr. Haixuan Yang
- Dr. Zenglin Xu (Ph.D.)
- Hongbo Deng (Ph.D.)
- Zhenjiang Lin (Ph.D.)
- Hao Ma (Ph.D.)
- Haiqin Yang (Ph.D.)
- Xin Xin (Ph.D.)
- Thomas Chan (M.Phil)
- Chao Zhou (Ph.D.)



VeriGuide

- **Similarity text** detection system
- Developed at **CUHK**
- Promote and uphold academic **honesty, integrity, and quality**
- Support **English, Traditional and Simplified Chinese**
- Handle **.doc, .txt, .pdf, .html**, etc. file formats
- Generate detailed **originality report** including **readability**



The screenshot shows the VeriGuide website interface. At the top, there is a navigation menu with links for HOME, SERVICES, NEWS, PARTNERS, ABOUT, REGISTER, and LOGIN. Below the menu is a large banner image of a university campus. The main content area is divided into three columns, each with a green header and a 'LEARN MORE' button:

- Originality:** VeriGuide aims to promote academic integrity with VeriGuide's similarity detection capabilities.
- Readability:** VeriGuide assists educators to assess students' writing ability over time with our readability features.
- Accessibility:** VeriGuide serves to support and maintain assignment database for students and educators.

Below these columns are three sections:

- VERIGUIDE:** A list of features including promoting academic honesty, supporting similarity text detection in multiple languages, providing class and assignment management, and maintaining assignment databases.
- WHAT'S NEW?:** News items about new service plans, the renaming of the Chinese University Plagiarism Identification Engine (CUPIDE) to VeriGuide, and the introduction of version 2.0 with improved processing and search capabilities.
- RESOURCES:** Links to Introduction, Tutorial - File submission, Tutorial - Report viewing, and Contact us.

The footer of the page includes the CUHK logo and a 'LEARN MORE' button.

<http://groups.google.com/group/WSCE2009>

Call for Papers



Workshop on Social Computing in Education (WSCE2009)
in conjunction with SocialComp-09, August 29-31, 2009, Vancouver, Canada

- Theory and modeling of social computing in education
- Technology and software of social computing for education
- Social educational system design and architectures
- Case studies, best practices, and demos of social media in education
- Benchmark and experiments on social computing in education
- Mobile learning applications for social computing
- Semantic web standards for e-learning
- Software for social learning and collaborative learning
- Life long social learning network
- Quality and reliability of information and resources
- Privacy, risk and security issues in education using social media
- Virtual space for leaning communities
- Ubiquitous, distributed, and collaborative learning
- Integration of social learning spaces
- Social computing in education trend analysis
- Web 2.0 and social computing for learning (media sharing, media manipulation, conversational arenas, online games, virtual worlds, social networking, blogging, social bookmarking, recommender systems, collaborative editing, wikis, syndication, etc.)



Q & A

<http://www.cse.cuhk.edu.hk/~king>

