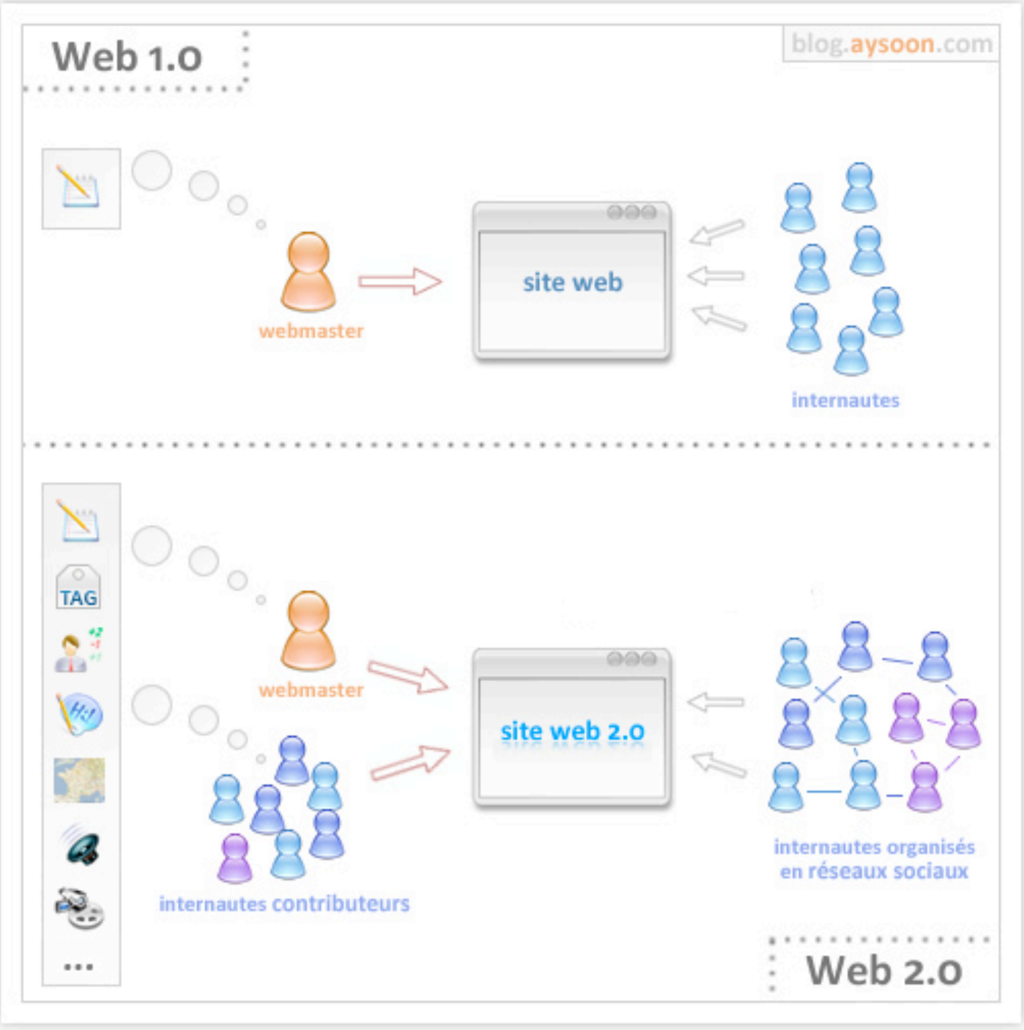
Social Network Analysis

bouguessa.mohamed@uqam.ca

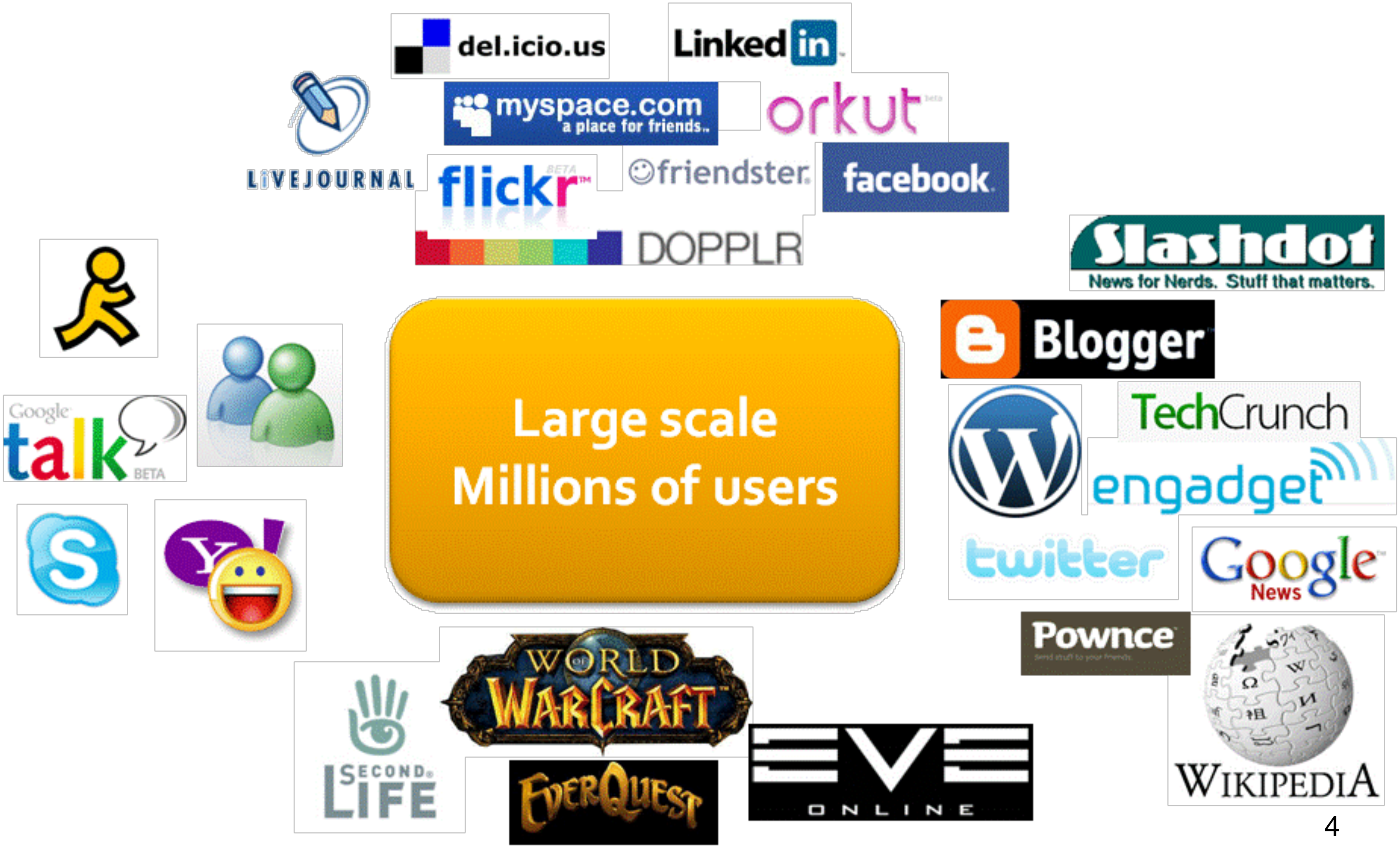
Web today


Web 1.0
blog.aysoon.com
webmaster
site web
internautes
TAG
webmaster
site web 2.0
internautes contributeurs
internautes organisés
en réseaux sociaux
Web 2.0

Web today – Diverse applications

Diverse on-line computing applications

Web today – Millions of users



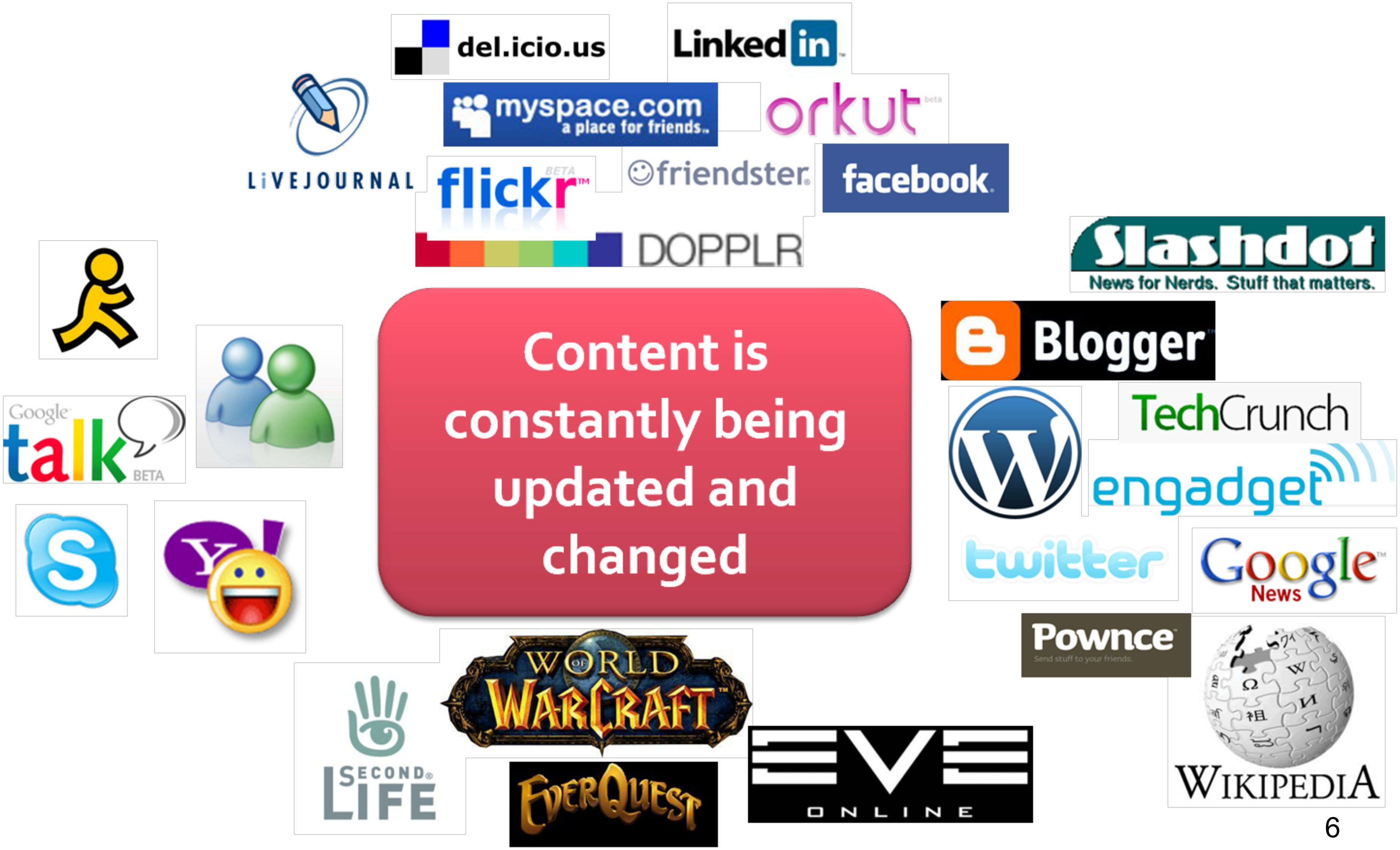
Large scale
Millions of users

Web today – Rich content

Rich user generated content

del.icio.us
LinkedIn
LiVEJOURNAL
myspace.com
a place for friends.
orkut
beta
flickr
BETA
friendster
facebook
DOPPLR
Slashdot
News for Nerds. Stuff that matters.
Blogger
TechCrunch
engadget
twitter
Google
News
Pownce
Send stuff to your friends.
Google
talk
BETA
Y!
SECOND LIFE
WORLD OF WARCRAFT
EVERQUEST
EVE
ONLINE
WIKIPEDIA

Web today – Highly dynamic



Content is constantly being updated and changed

Web today – Traces of activity

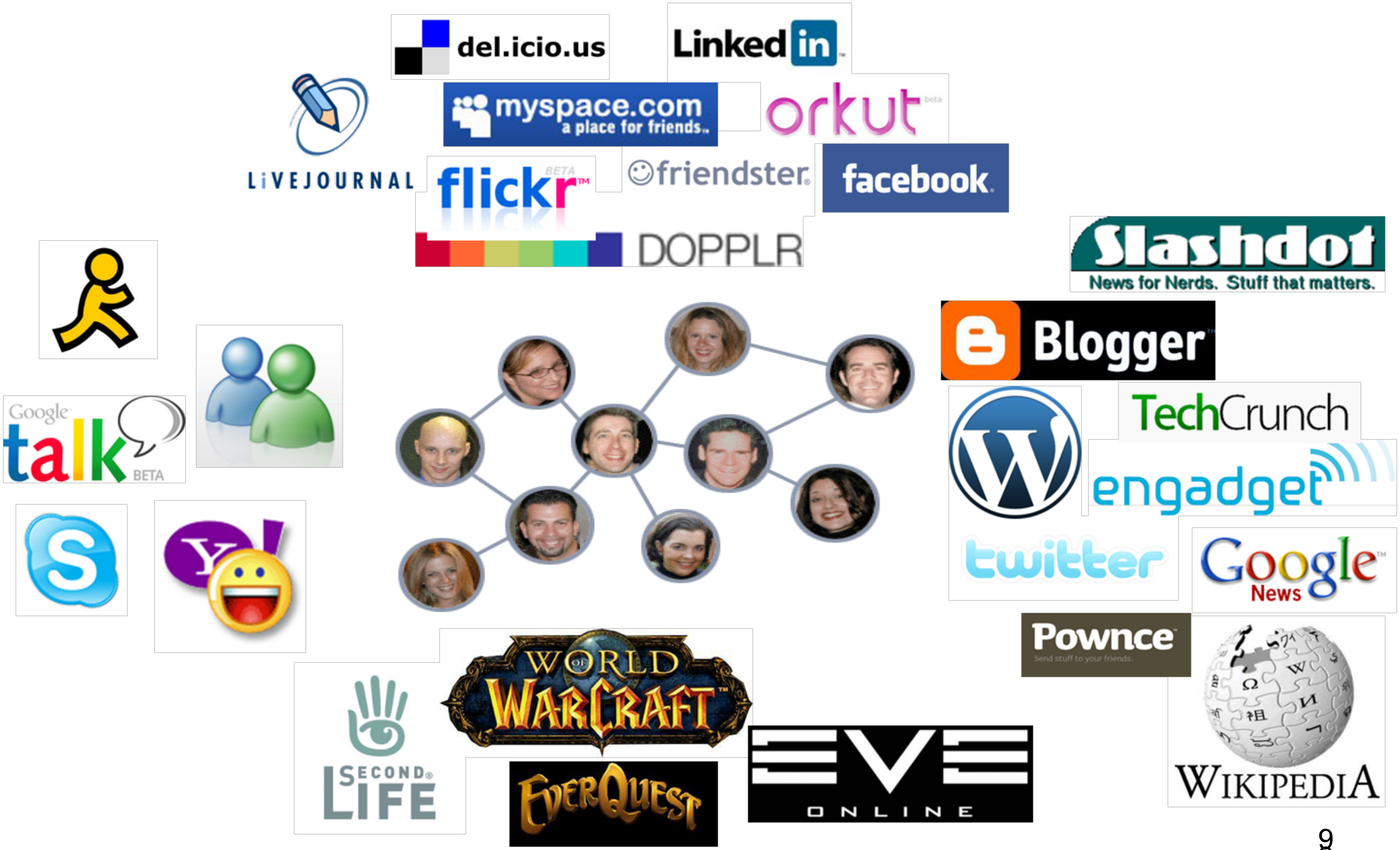
Massive traces of human social activity are collected

Web today – Rich interactions

Rich interactions between users and content

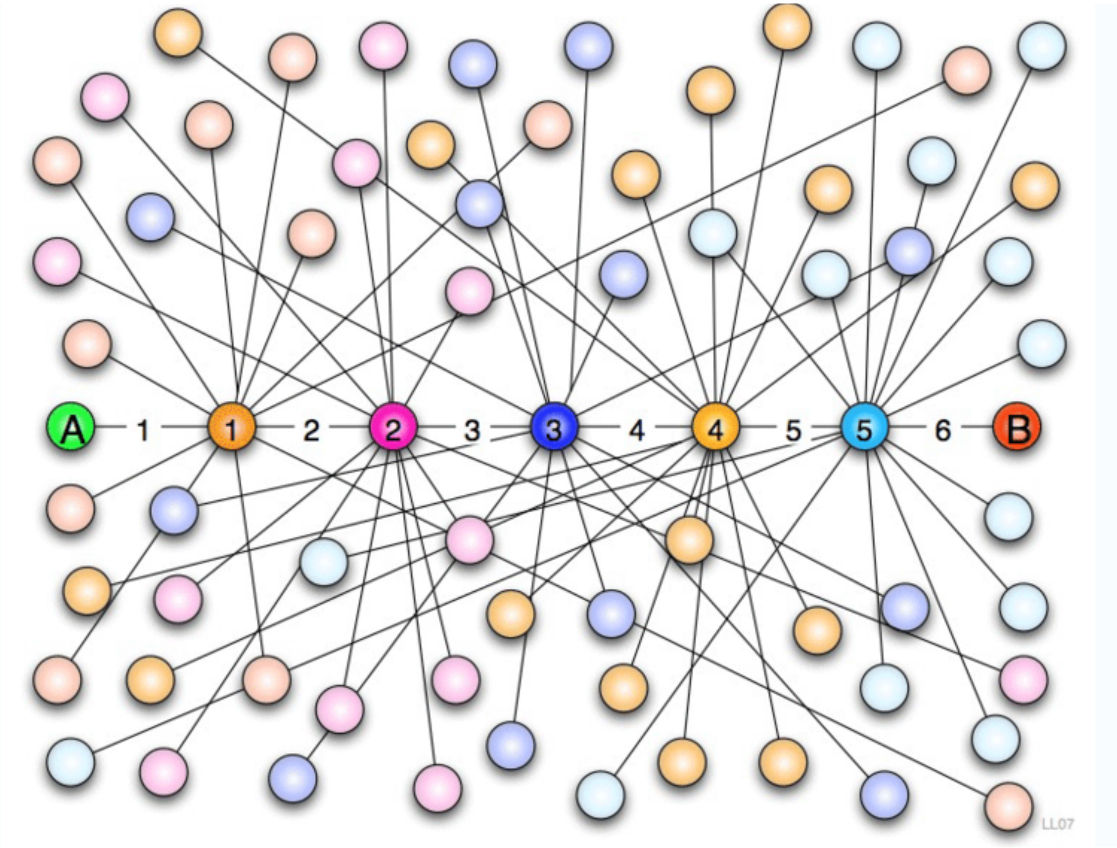
del.icio.us
LinkedIn
LiVEJOURNAL
myspace.com a place for friends..
orkut beta
flickr BETA
friendster
facebook
DOPPLR
Slashdot News for Nerds. Stuff that matters.
Google talk BETA
Blogger
TechCrunch
engadget
twitter
Google News
Pownce Send stuff to your friends.
WORLD OF WARCRAFT
SECOND LIFE
EVERQUEST
EVE ONLINE
WIKIPEDIA

Web today – social networks



Six degrees of separation

We can all be connected through a series of six contacts appeals to me. It makes the world seem less brutal, and more warm and more friendly.



Why study networks?

- **Build understanding and theory:**
  - How users create content and interact with it and among themselves?

- **Build better on-line applications:**
  - How to design better services and algorithms?

Social Networks Analysis

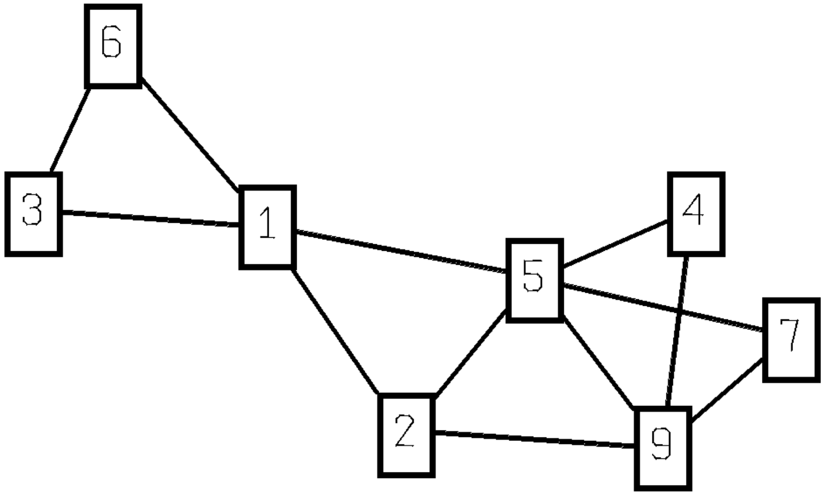
- A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest.

- **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior.

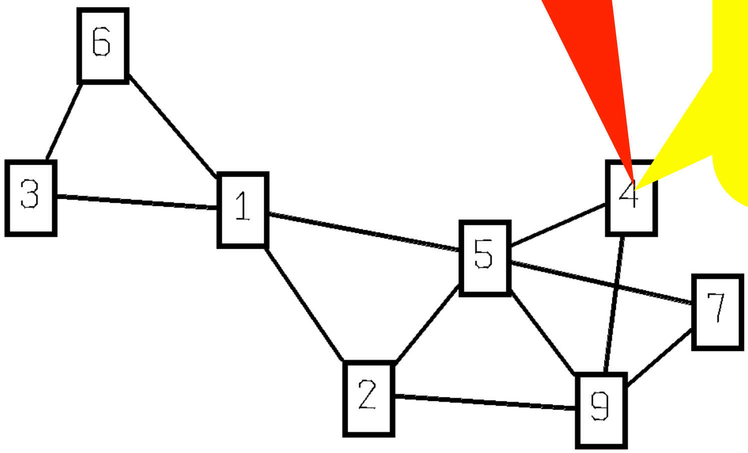
WELL, IF A LIKES B, BUT B LIKES C WHO LIKES D AND E WHO BOTH LIKE A WHO DOESN'T EVEN KNOW THAT D EXISTS, SHOULD F TRY TO HAVE G TALK TO B SO E WILL KNOW THAT C LIKES D AND E, AND THAT C WILL POUND H IF SHE COMES AROUND AGAIN BUTTING IN?
MAY I THINK ABOUT THAT FOR A MINUTE?

Social Networks

- Social network: relationship among interacting units.



Social Networks



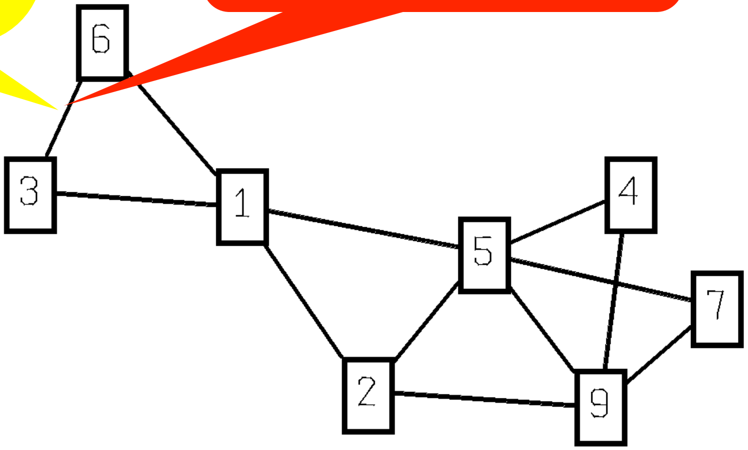
Interacting unites:
Actors / nodes

discrete individual,
corporate, or
collective social units

Social Networks

Relational ties between actors are channels to transfer, exchange or flow of resources.

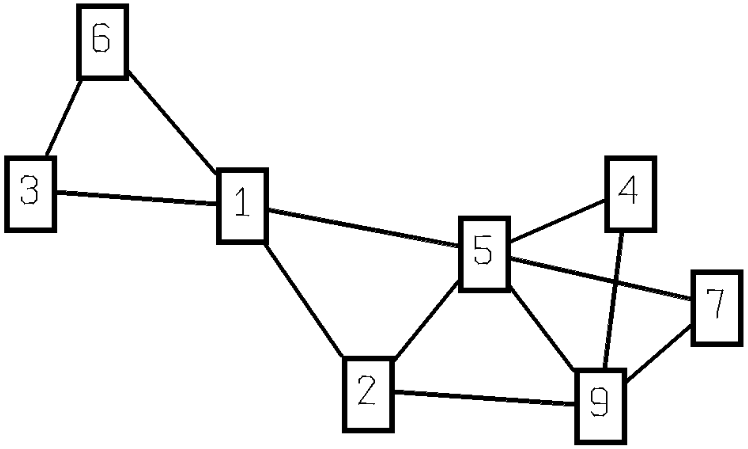
Relations, linkages or ties



Social Networks

- Social network representation
  - Adjacency matrix (socio-matrix)
  - Graph (Socio-graph)

$$\begin{array}{c|ccccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline
1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\
2 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
3 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
5 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\
6 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
7 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
9 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0
\end{array}$$



Key Drivers for CS Research in SNA

- Computer Science has created the cyber infrastructure for
  - Social Interaction
  - Knowledge Exchange
  - Knowledge Discovery

- Ability to capture
  - different about various types of social interactions
  - at a very fine granularity
  - with practically no reporting bias

Data mining techniques can be used for building descriptive and predictive models of social interactions

SNA Techniques

Prominent problems

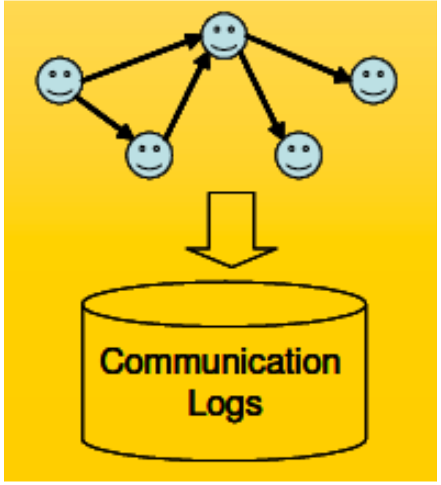
- Social network extraction/construction
- Identifying prominent/trusted/expert actors
- Identifying Spammers
- Discovering communities in social networks
- Evolution of social networks
- Link prediction
- Approximating large social networks

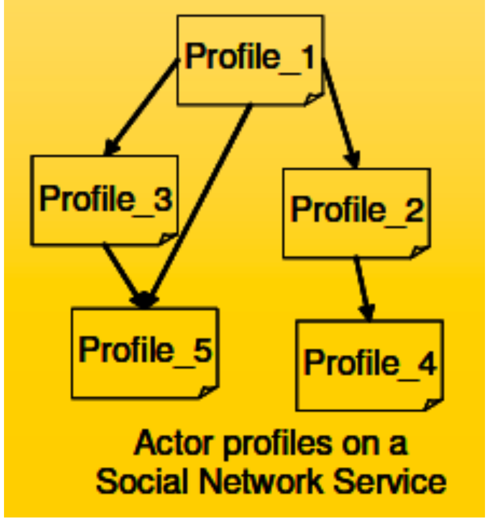
Social Network Extraction

- Mining a social network from data sources
- Recent research suggest that there are three sources of social network data on the web

- Content available on web pages (e.g. user homepages, message threads etc.)
- User interaction logs (e.g. email and messenger chat logs)
- Social interaction information provided by users (e.g. social network service websites such as Orkut, Friendster and MySpace)

Web Documents



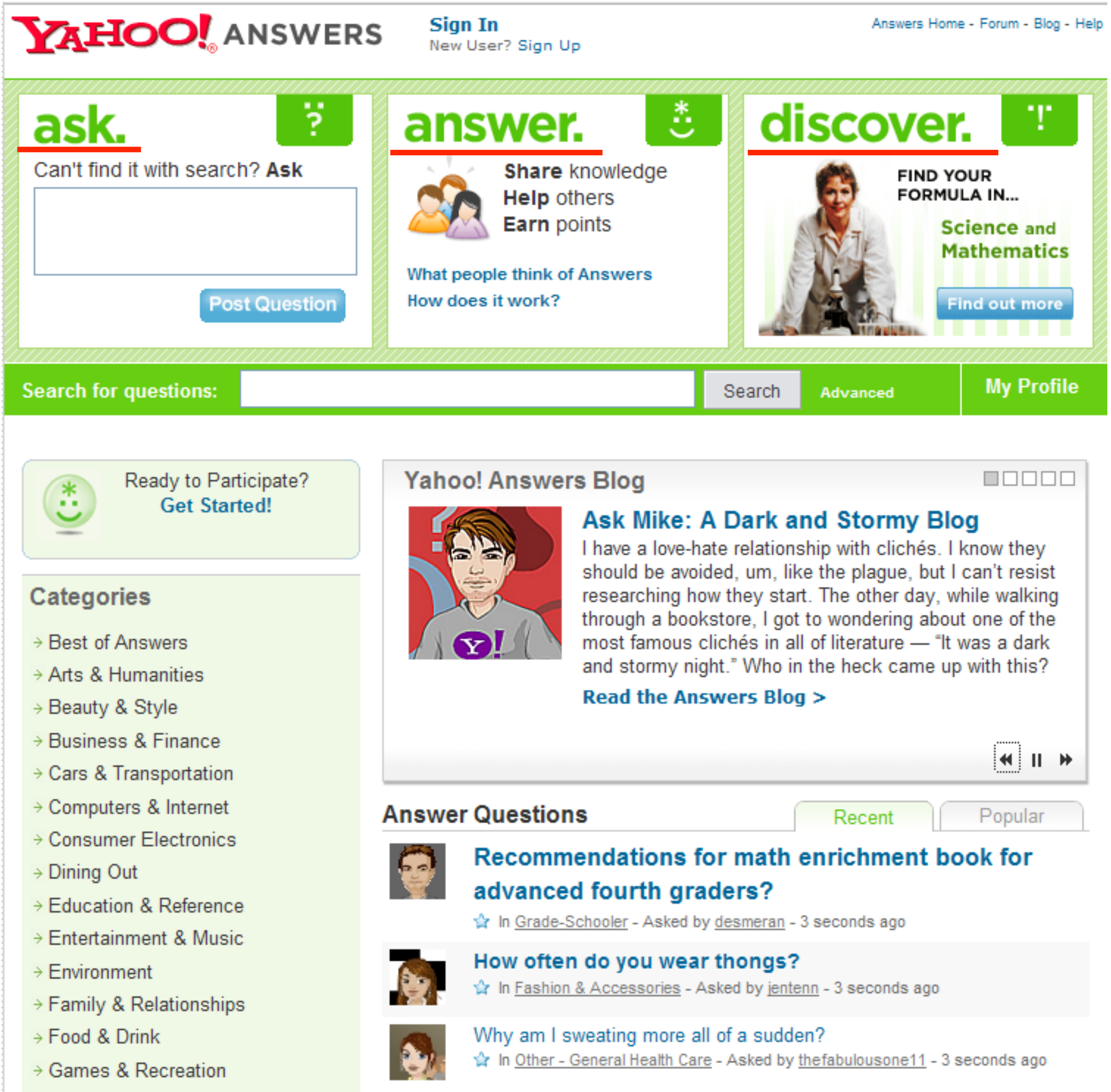


SNA Techniques

Prominent problems

- Social network extraction/construction
- **Identifying prominent/trusted/expert actors**
- Identifying Spammers
- Discovering communities in social networks
- Link prediction
- Approximating large social networks
- Evolution of social networks

Yahoo! Answers



YAHOO! ANSWERS

Sign In
New User? Sign Up

Answers Home - Forum - Blog - Help

ask.
Can't find it with search? Ask
Post Question

answer.
Share knowledge
Help others
Earn points

What people think of Answers
How does it work?

discover.
FIND YOUR FORMULA IN...
Science and Mathematics
Find out more

Search for questions: Search Advanced

My Profile

Ready to Participate?
Get Started!

Categories
- Best of Answers
- Arts & Humanities
- Beauty & Style
- Business & Finance
- Cars & Transportation
- Computers & Internet
- Consumer Electronics
- Dining Out
- Education & Reference
- Entertainment & Music
- Environment
- Family & Relationships
- Food & Drink
- Games & Recreation

Yahoo! Answers Blog

Ask Mike: A Dark and Stormy Blog

I have a love-hate relationship with clichés. I know they should be avoided, um, like the plague, but I can't resist researching how they start. The other day, while walking through a bookstore, I got to wondering about one of the most famous clichés in all of literature — "It was a dark and stormy night." Who in the heck came up with this?

Read the Answers Blog >

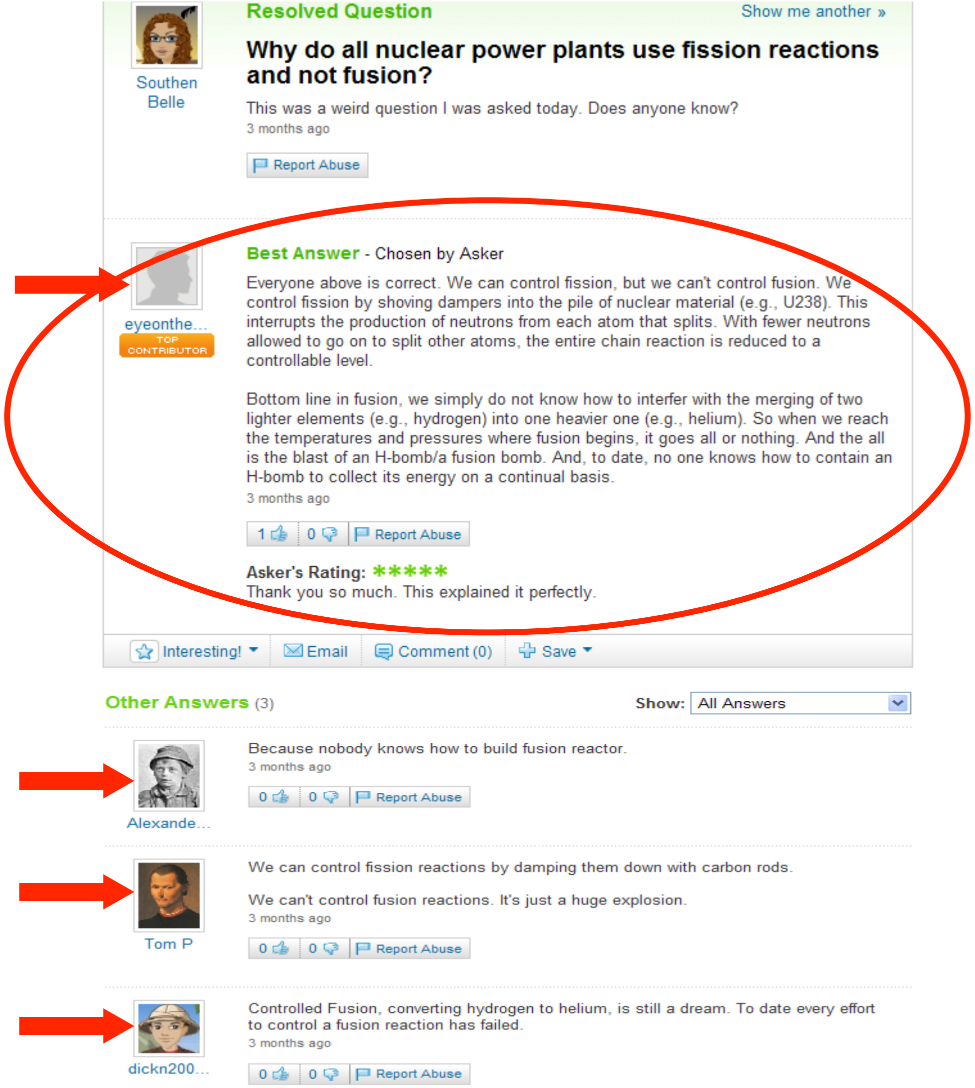
Answer Questions
Recent Popular

Recommendations for math enrichment book for advanced fourth graders?
In Grade-Schooler - Asked by desmeran - 3 seconds ago

How often do you wear thongs?
In Fashion & Accessories - Asked by jentenn - 3 seconds ago

Why am I sweating more all of a sudden?
In Other - General Health Care - Asked by thefabulousone11 - 3 seconds ago

Question Life Cycle



Resolved Question

Show me another »

Southen Belle

Why do all nuclear power plants use fission reactions and not fusion?

This was a weird question I was asked today. Does anyone know?

3 months ago

Report Abuse

eyeonthe... TOP CONTRIBUTOR

Best Answer - Chosen by Asker

Everyone above is correct. We can control fission, but we can't control fusion. We control fission by shoving dampers into the pile of nuclear material (e.g., U238). This interrupts the production of neutrons from each atom that splits. With fewer neutrons allowed to go on to split other atoms, the entire chain reaction is reduced to a controllable level.

Bottom line in fusion, we simply do not know how to interfer with the merging of two lighter elements (e.g., hydrogen) into one heavier one (e.g., helium). So when we reach the temperatures and pressures where fusion begins, it goes all or nothing. And the all is the blast of an H-bomb/a fusion bomb. And, to date, no one knows how to contain an H-bomb to collect its energy on a continual basis.

3 months ago

1 | 0 | Report Abuse

Asker's Rating: *****

Thank you so much. This explained it perfectly.

Interesting! | Email | Comment (0) | Save

Other Answers (3)

Show: All Answers

Alexande...

Because nobody knows how to build fusion reactor.

3 months ago

0 | 0 | Report Abuse

Tom P

We can control fission reactions by damping them down with carbon rods.

We can't control fusion reactions. It's just a huge explosion.

3 months ago

0 | 0 | Report Abuse

dickn200...

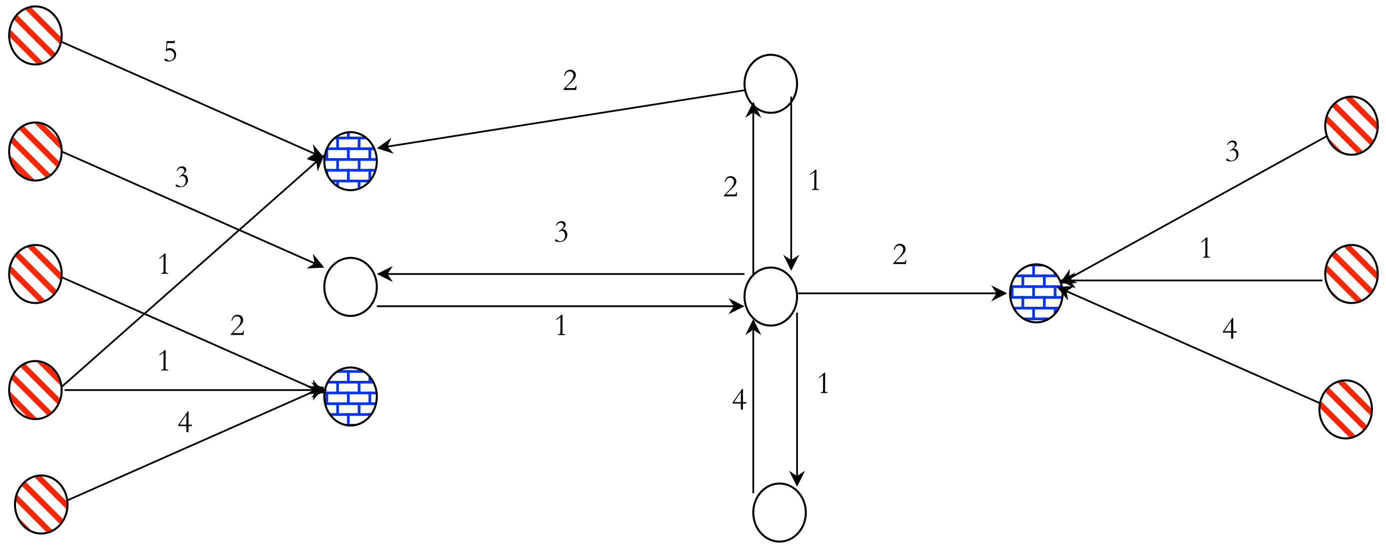
Controlled Fusion, converting hydrogen to helium, is still a dream. To date every effort to control a fusion reaction has failed.

3 months ago

0 | 0 | Report Abuse

Yahoo! Answers

Example of interactions between askers and best answerers



▨ Users who usually only ask questions

▦ Users who usually only answer questions

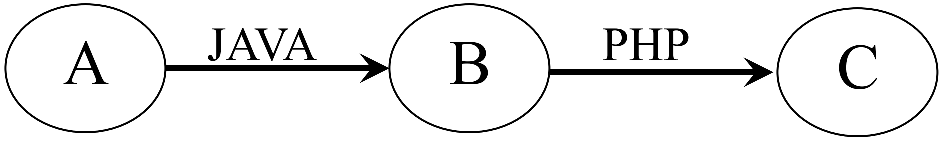
○ Users who help each other

How to estimate the authority degree for each user?

PageRank?

Example: The category of “Programming”

- User *B* answers user *A*’s questions, which are about Java;
- User *C* answers *B*’s questions, which are about PHP;



➢ Is it possible to state that *C* is more expert than *B*?

- No, because: *B* and *C* have different expertise.

Proposed Approach

- The authority score of each user is simply the number of best answer of each users normalized so their square sum to 1:

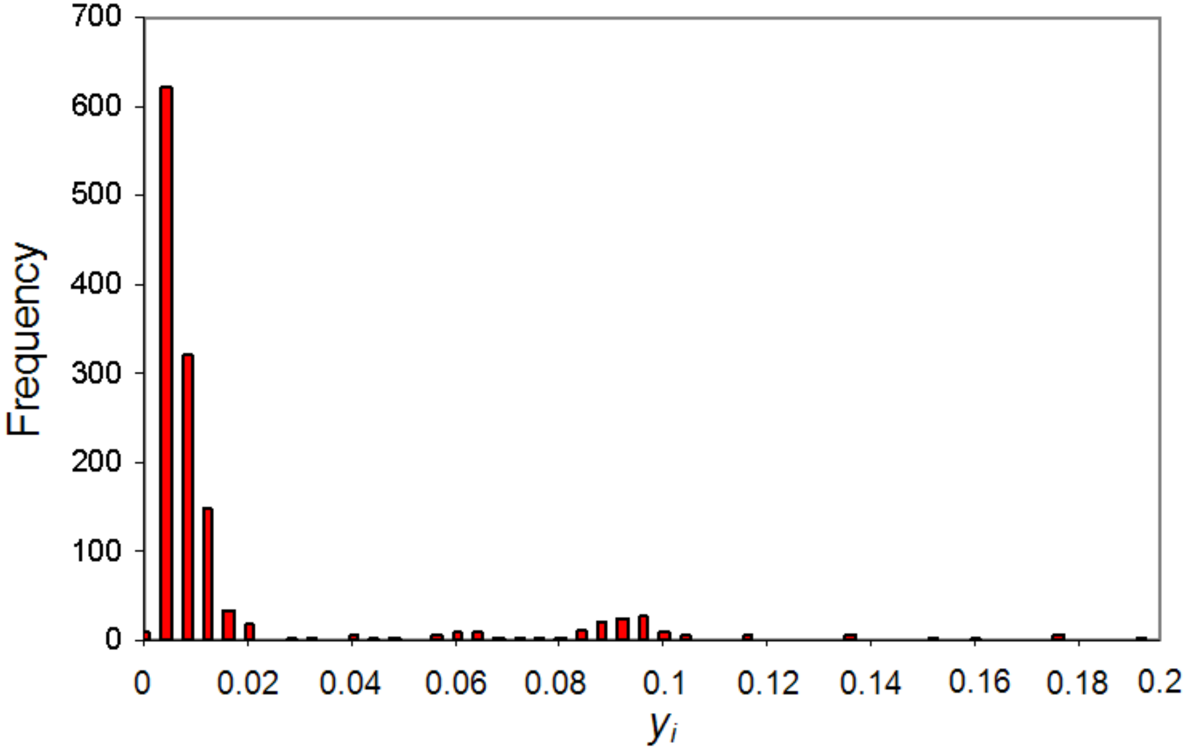
$$\sum_{i=1}^{N} (y_i)^2 = 1$$

- y_i provide a relative score of the authority of each user in each category.

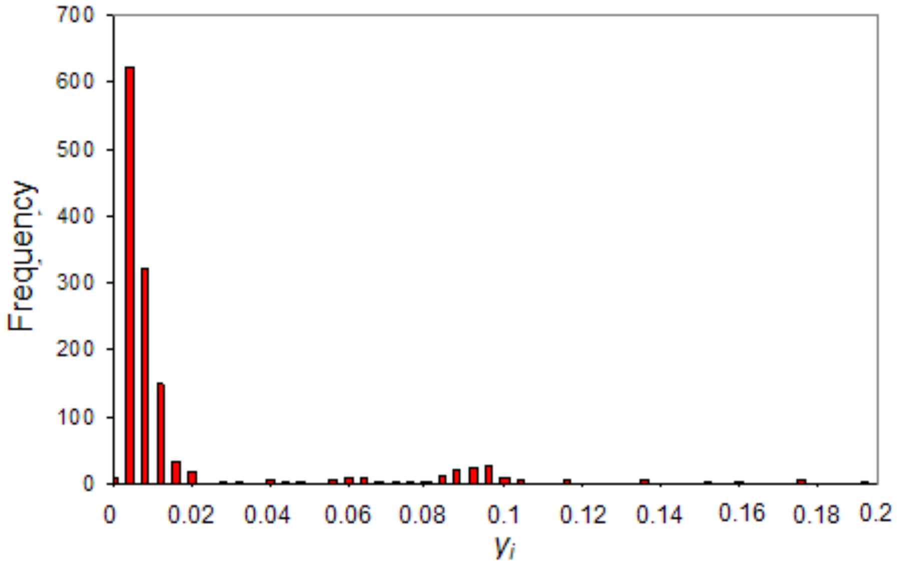
➢ We are interested in all sets of U_i having large values of y_i.

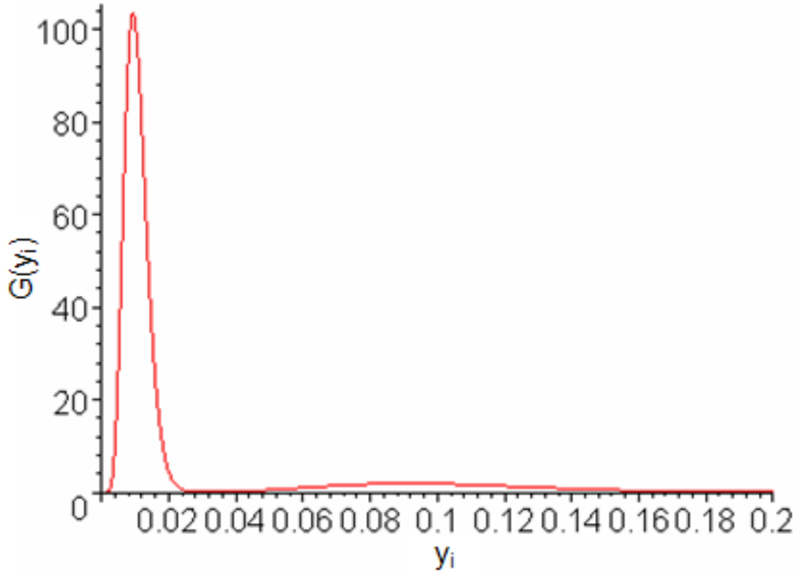
Authority Score

- Example: Category of "Engineering"

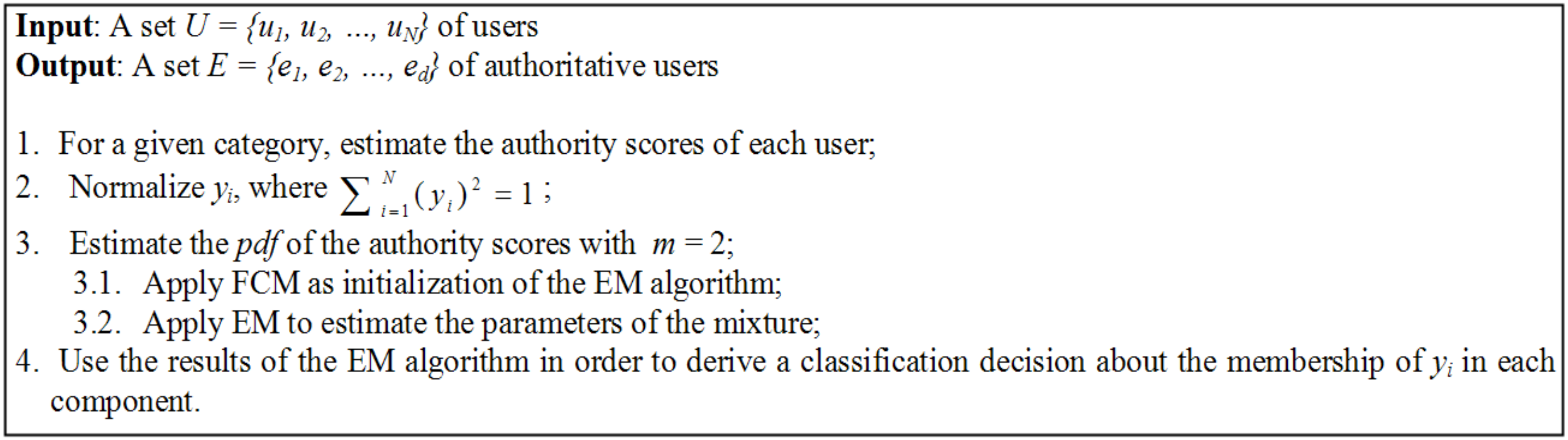


Authority Score





Automatic Identification of Authorities



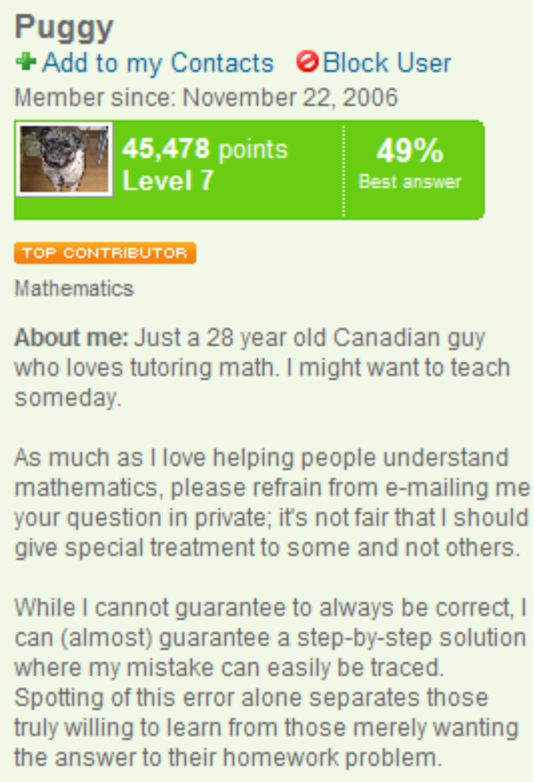
```
**Input**: A set $U = \{u_1, u_2, ..., u_N\}$ of users
**Output**: A set $E = \{e_1, e_2, ..., e_d\}$ of authoritative users

1. For a given category, estimate the authority scores of each user;
2. Normalize $y_i$, where $\sum_{i=1}^{N} (y_i)^2 = 1$;
3. Estimate the *pdf* of the authority scores with $m = 2$;
    3.1. Apply FCM as initialization of the EM algorithm;
    3.2. Apply EM to estimate the parameters of the mixture;
4. Use the results of the EM algorithm in order to derive a classification decision about the membership of $y_i$ in each
   component.
```

Experiments

We conduct experiments on datasets which represent users' activities over one full year for six categories:

Category	% users who ask only	% users who answer only	% users who ask and answer
Engineering	65%	31%	4%
Biology	60%	36%	4%
Programming	66%	29%	5%
Mathematics	64%	31%	5%
Physics	60%	34%	6%
Chemistry	63%	32%	5%

Authoritative Users



Puggy

Member since: November 22, 2006

45,478 points
Level 7

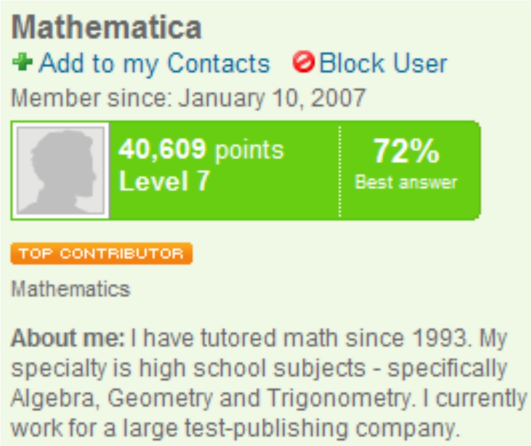
49%
Best answer

Mathematics

About me: Just a 28 year old Canadian guy who loves tutoring math. I might want to teach someday.

As much as I love helping people understand mathematics, please refrain from e-mailing me your question in private; it's not fair that I should give special treatment to some and not others.

While I cannot guarantee to always be correct, I can (almost) guarantee a step-by-step solution where my mistake can easily be traced. Spotting of this error alone separates those truly willing to learn from those merely wanting the answer to their homework problem.



Mathematica

Member since: January 10, 2007

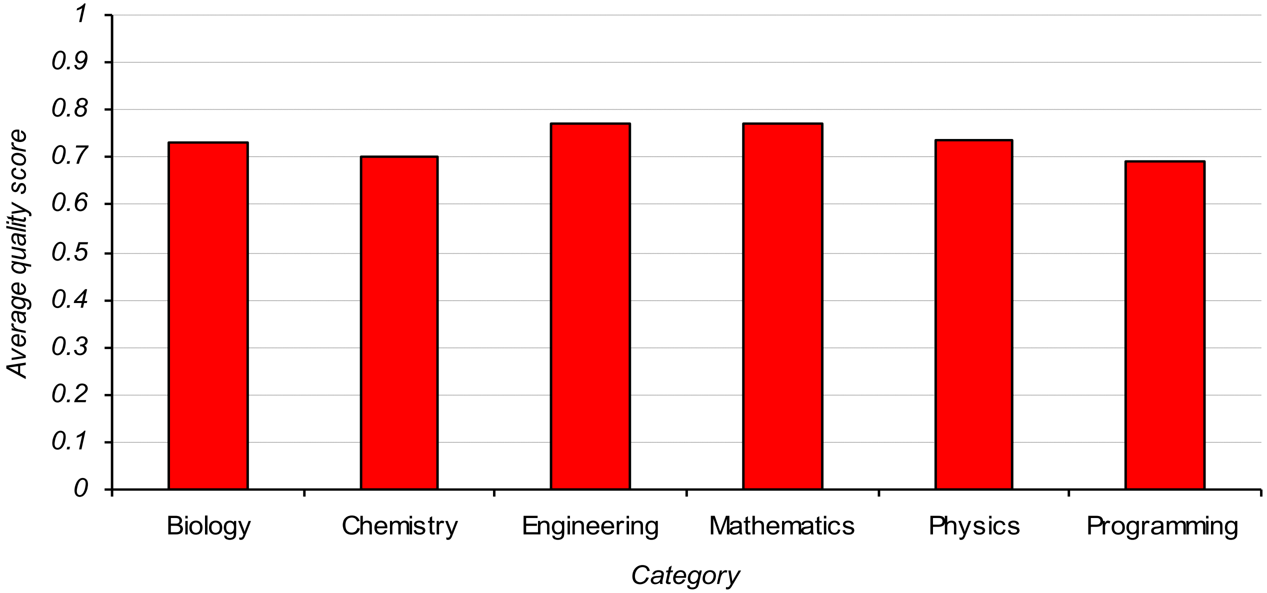
40,609 points
Level 7

72%
Best answer

Mathematics

About me: I have tutored math since 1993. My specialty is high school subjects - specifically Algebra, Geometry and Trigonometry. I currently work for a large test-publishing company.

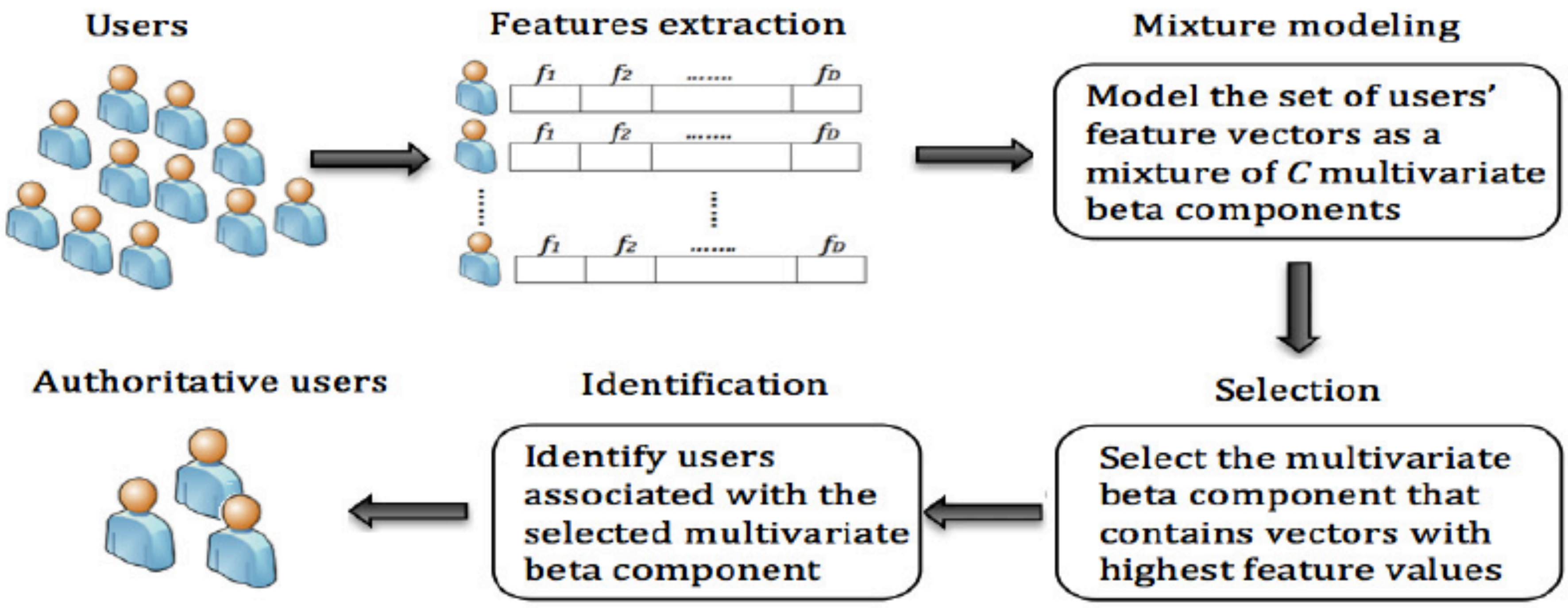
Quality of Content



- The identified authoritative users generate high-quality content in Yahoo! Answers.

- Askers are very selective in choosing the best answerers

Identifying Authorities in Online Communities


Users
Features extraction
f_1 f_2 .... f_D
f_1 f_2 .... f_D
f_1 f_2 .... f_D
Mixture modeling
Model the set of users' feature vectors as a mixture of *C* multivariate beta components
Selection
Select the multivariate beta component that contains vectors with highest feature values
Identification
Identify users associated with the selected multivariate beta component
Authoritative users

Workflow of the proposed approach.

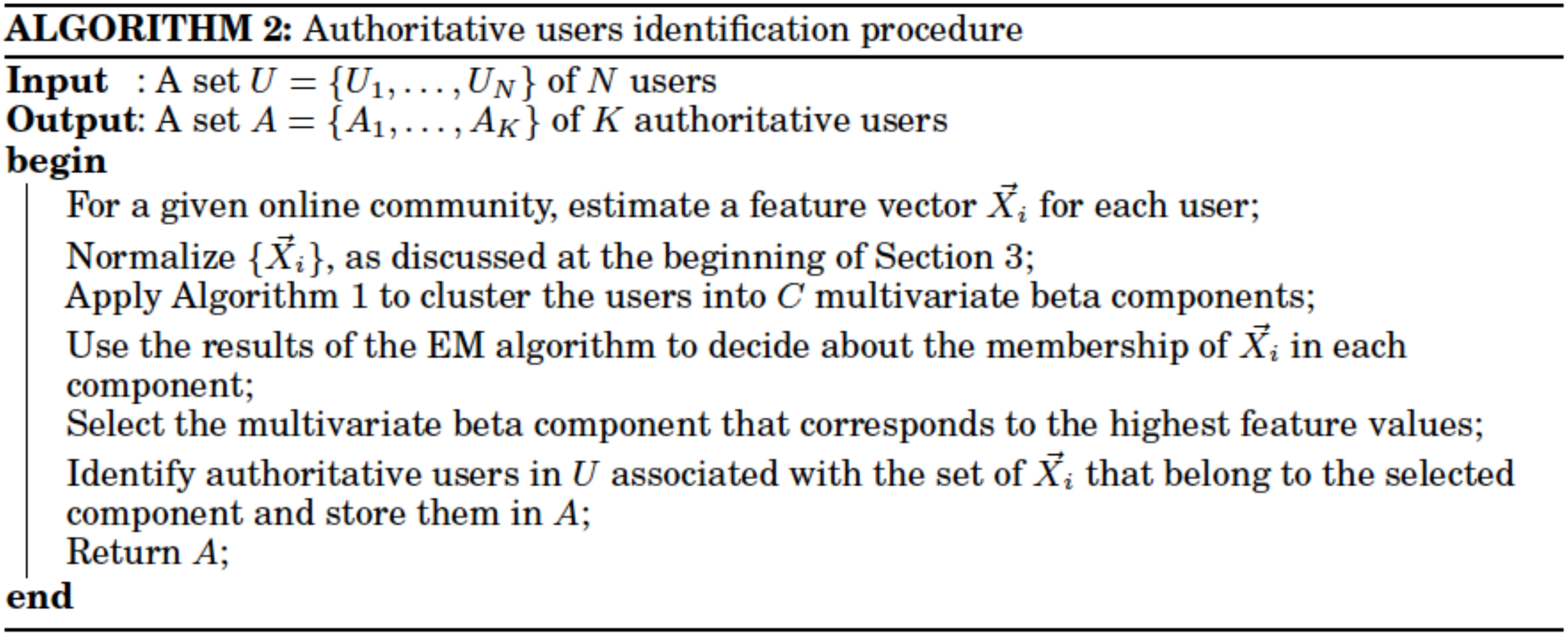
Multivariate Beta Mixture Model

$$\mathcal{F}(\vec{X}_i|\alpha,\vec{a},\vec{b}) = \sum_{c=1}^{C} \alpha_c\, \mathcal{F}_c(\vec{X}_i|\vec{a_c},\vec{b_c})$$

$$\mathcal{F}_c(\vec{X}_i|\vec{a_c},\vec{b_c}) = \prod_{d=1}^{D} f(x_{id}|a_{cd},b_{cd})$$

$$f(x_{id}|a_{cd},b_{cd}) = \frac{\Gamma(a_{cd}+b_{cd})}{\Gamma(a_{cd})\Gamma(b_{cd})} x_{id}^{a_{cd}-1}(1-x_{id})^{b_{cd}-1}$$

Algorithm



ALGORITHM 2: Authoritative users identification procedure

```
Input  : A set $U = \{U_1, \dots, U_N\}$ of $N$ users
Output: A set $A = \{A_1, \dots, A_K\}$ of $K$ authoritative users
begin
    For a given online community, estimate a feature vector $\vec{X}_i$ for each user;
    Normalize $\{\vec{X}_i\}$, as discussed at the beginning of Section 3;
    Apply Algorithm 1 to cluster the users into $C$ multivariate beta components;
    Use the results of the EM algorithm to decide about the membership of $\vec{X}_i$ in each
    component;
    Select the multivariate beta component that corresponds to the highest feature values;
    Identify authoritative users in $U$ associated with the set of $\vec{X}_i$ that belong to the selected
    component and store them in $A$;
    Return $A$;
end
```

Twitter data – 2012 Quebec election

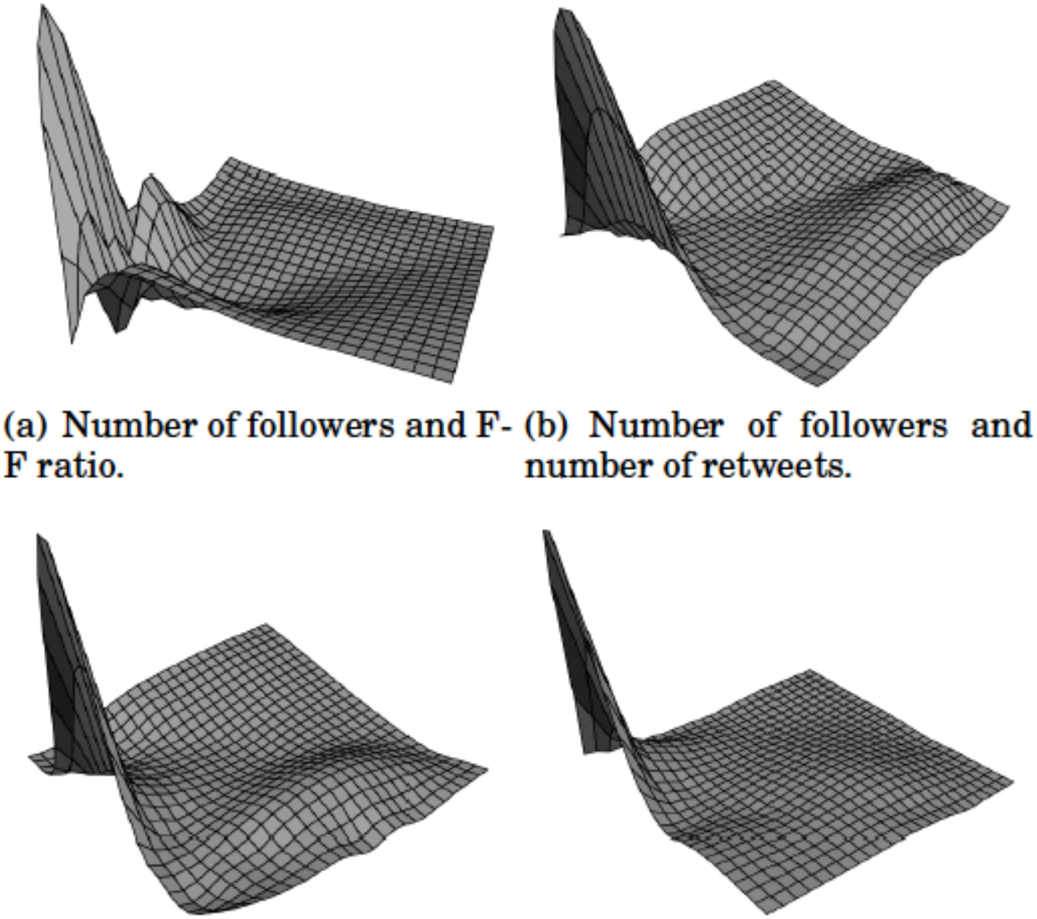
- The data set consists of tweets posted between August 18, 2012 and August 20, 2012 (three days overall during the electoral campaign, including Quebec's political party leaders' debate which took place on August 19, 2012).

- 904 users; 76 users (8.4% of the whole data set) among them were labeled as authoritative and 828 users were labelled as non-authoritative

Twitter data – 2012 Quebec election

- Features
  - The number of followers of a user, which indicates the size of the audience for that user
  - The Followers to Followees ratio (F-F ratio), that is, the number of a user’s followers and the number of other people that the user follows (followees).
  - The number of retweets, which measures the number of times an author’s tweets were retweeted by other users
  - The number of mentions, which is measured by the number of times a user was cited or had her tweet replied to.

Twitter data – 2012 Quebec election



(a) Number of followers and F-F ratio.

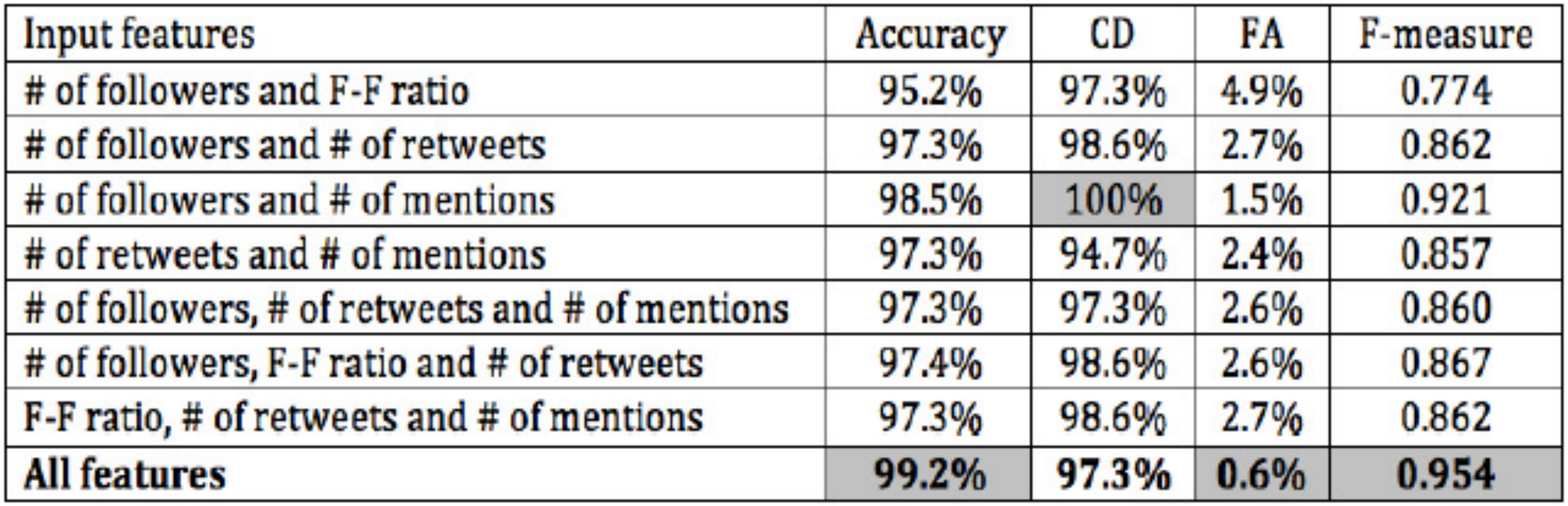
(b) Number of followers and number of retweets.

(c) Number of followers and number of mentions.

(d) Number of retweets and number of mentions.

Density curves of several $2D$ user features combinations over Quebec Election

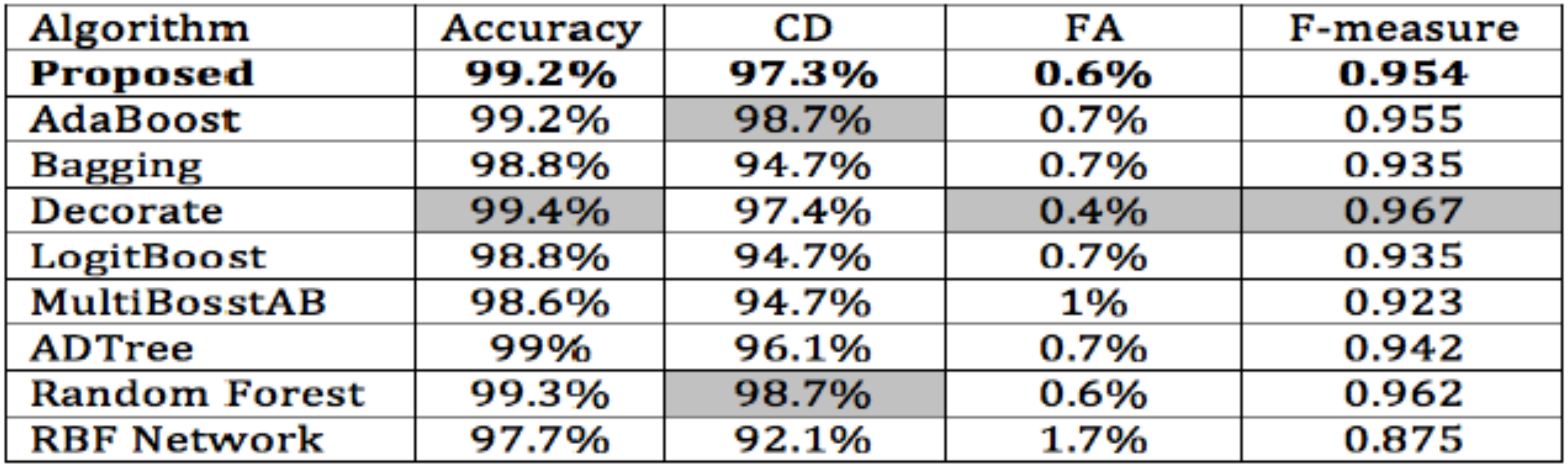
Twitter data – 2012 Quebec election



Input features	Accuracy	CD	FA	F-measure
# of followers and F-F ratio	95.2%	97.3%	4.9%	0.774
# of followers and # of retweets	97.3%	98.6%	2.7%	0.862
# of followers and # of mentions	98.5%	100%	1.5%	0.921
# of retweets and # of mentions	97.3%	94.7%	2.4%	0.857
# of followers, # of retweets and # of mentions	97.3%	97.3%	2.6%	0.860
# of followers, F-F ratio and # of retweets	97.4%	98.6%	2.6%	0.867
F-F ratio, # of retweets and # of mentions	97.3%	98.6%	2.7%	0.862
All features	**99.2%**	**97.3%**	**0.6%**	**0.954**

Performance results over Quebec Election data.

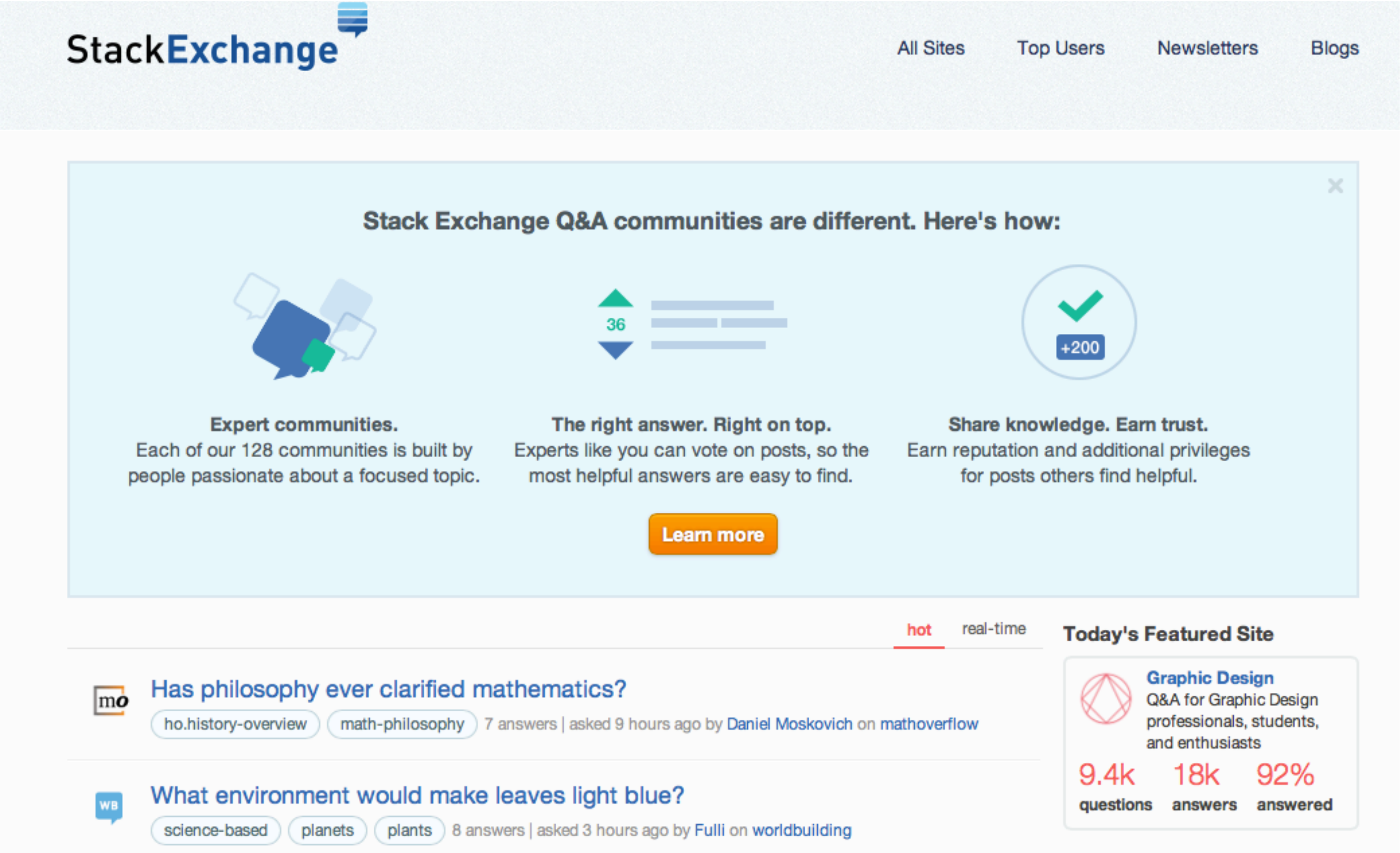
Twitter data – 2012 Quebec election



Algorithm	Accuracy	CD	FA	F-measure
Proposed	**99.2%**	**97.3%**	**0.6%**	**0.954**
AdaBoost	99.2%	98.7%	0.7%	0.955
Bagging	98.8%	94.7%	0.7%	0.935
Decorate	99.4%	97.4%	0.4%	0.967
LogitBoost	98.8%	94.7%	0.7%	0.935
MultiBosstAB	98.6%	94.7%	1%	0.923
ADTree	99%	96.1%	0.7%	0.942
Random Forest	99.3%	98.7%	0.6%	0.962
RBF Network	97.7%	92.1%	1.7%	0.875

Accuracies of compared algorithms on Quebec Election data.

Stack Exchange data



Stack Exchange Q&A communities are different. Here's how:

Expert communities.
Each of our 128 communities is built by people passionate about a focused topic.

The right answer. Right on top.
Experts like you can vote on posts, so the most helpful answers are easy to find.

Share knowledge. Earn trust.
Earn reputation and additional privileges for posts others find helpful.

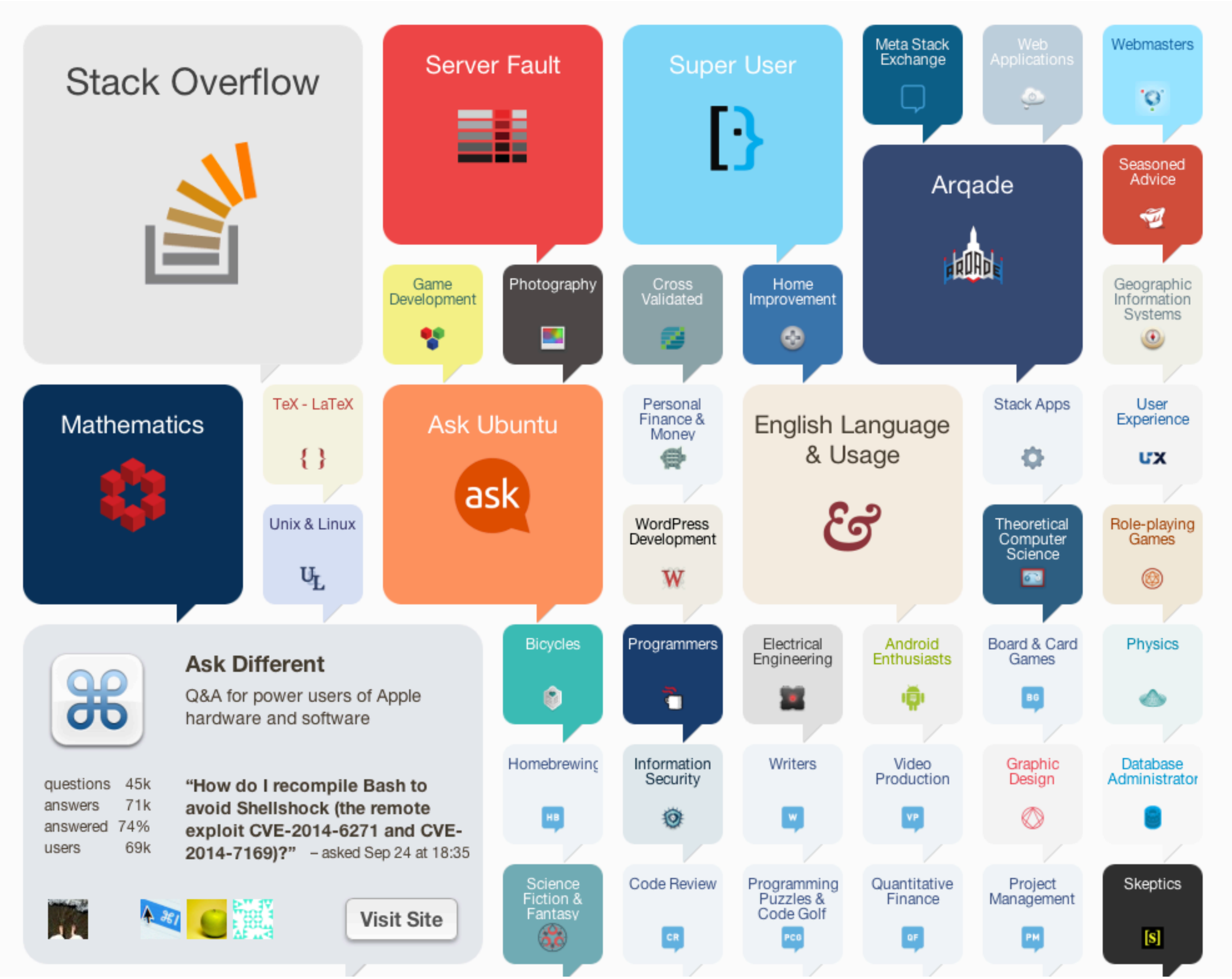
Has philosophy ever clarified mathematics?
ho.history-overview | math-philosophy | 7 answers | asked 9 hours ago by Daniel Moskovich on mathoverflow

What environment would make leaves light blue?
science-based | planets | plants | 8 answers | asked 3 hours ago by Fulli on worldbuilding

Today's Featured Site

Graphic Design
Q&A for Graphic Design professionals, students, and enthusiasts

9.4k	18k	92%
questions	answers	answered

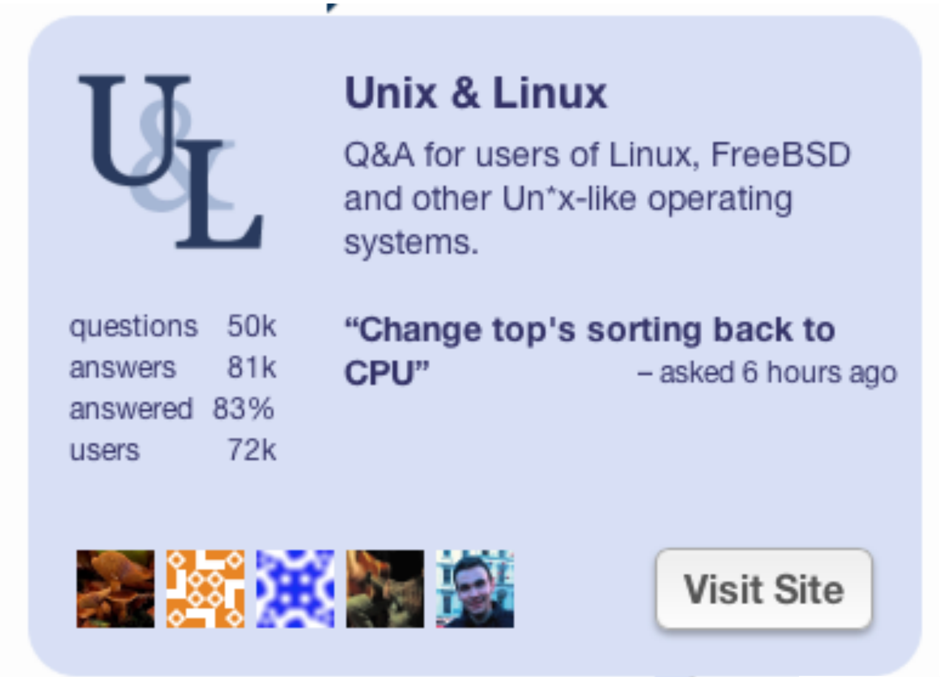


Ask Different

Q&A for power users of Apple hardware and software

questions	45k
answers	71k
answered	74%
users	69k

"How do I recompile Bash to avoid Shellshock (the remote exploit CVE-2014-6271 and CVE-2014-7169)?" – asked Sep 24 at 18:35

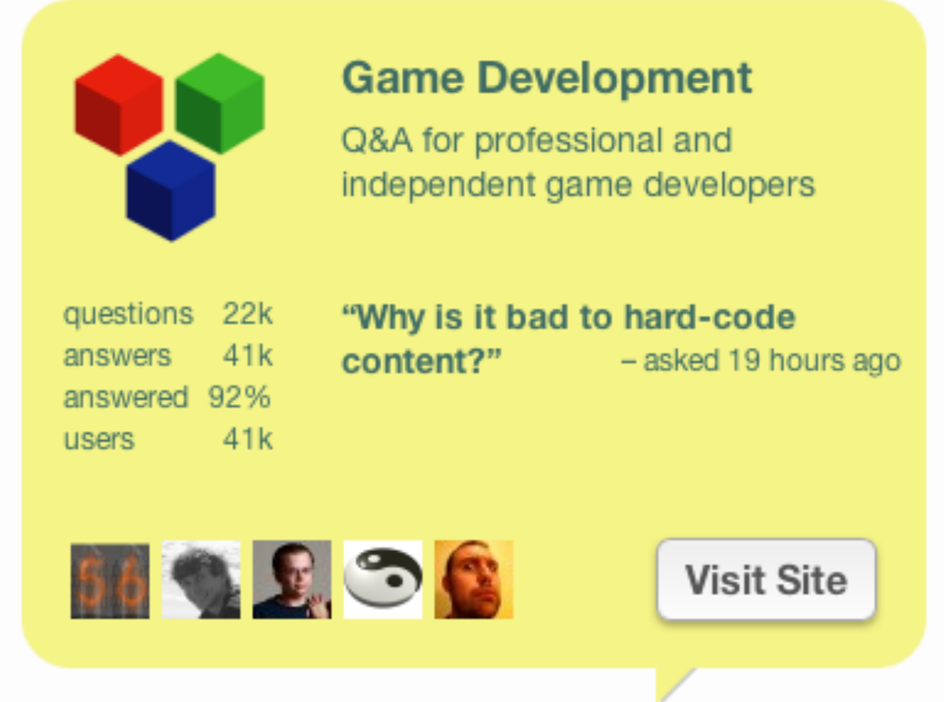


Unix & Linux

Q&A for users of Linux, FreeBSD and other Un*x-like operating systems.

questions	50k
answers	81k
answered	83%
users	72k

“Change top's sorting back to CPU” – asked 6 hours ago

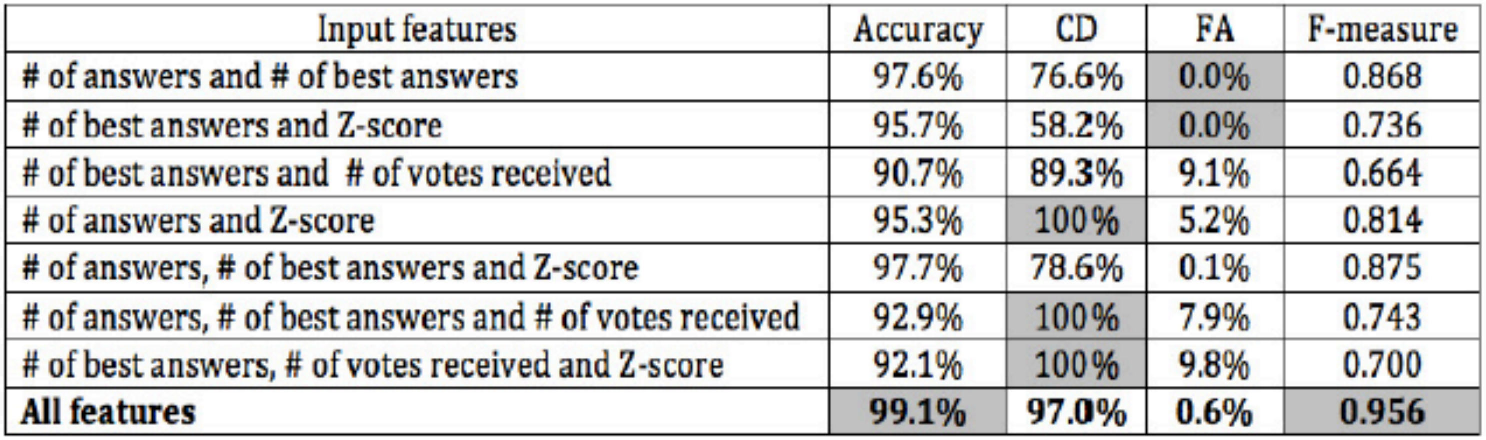


Game Development

Q&A for professional and independent game developers

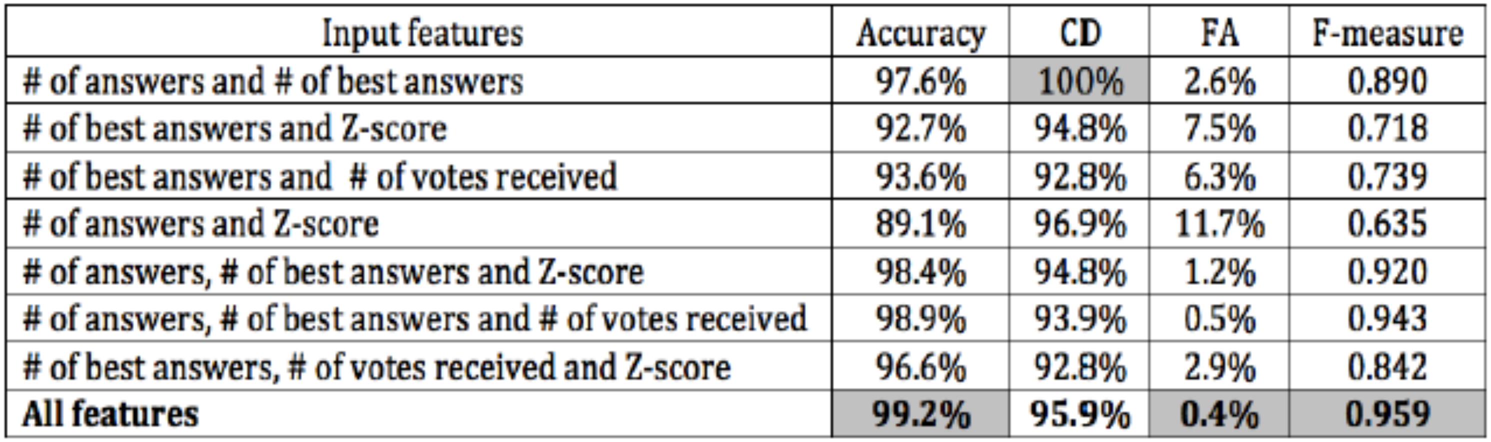
questions	22k
answers	41k
answered	92%
users	41k

“Why is it bad to hard-code content?” – asked 19 hours ago



Input features	Accuracy	CD	FA	F-measure
# of answers and # of best answers	97.6%	76.6%	0.0%	0.868
# of best answers and Z-score	95.7%	58.2%	0.0%	0.736
# of best answers and # of votes received	90.7%	89.3%	9.1%	0.664
# of answers and Z-score	95.3%	100%	5.2%	0.814
# of answers, # of best answers and Z-score	97.7%	78.6%	0.1%	0.875
# of answers, # of best answers and # of votes received	92.9%	100%	7.9%	0.743
# of best answers, # of votes received and Z-score	92.1%	100%	9.8%	0.700
All features	**99.1%**	**97.0%**	**0.6%**	**0.956**

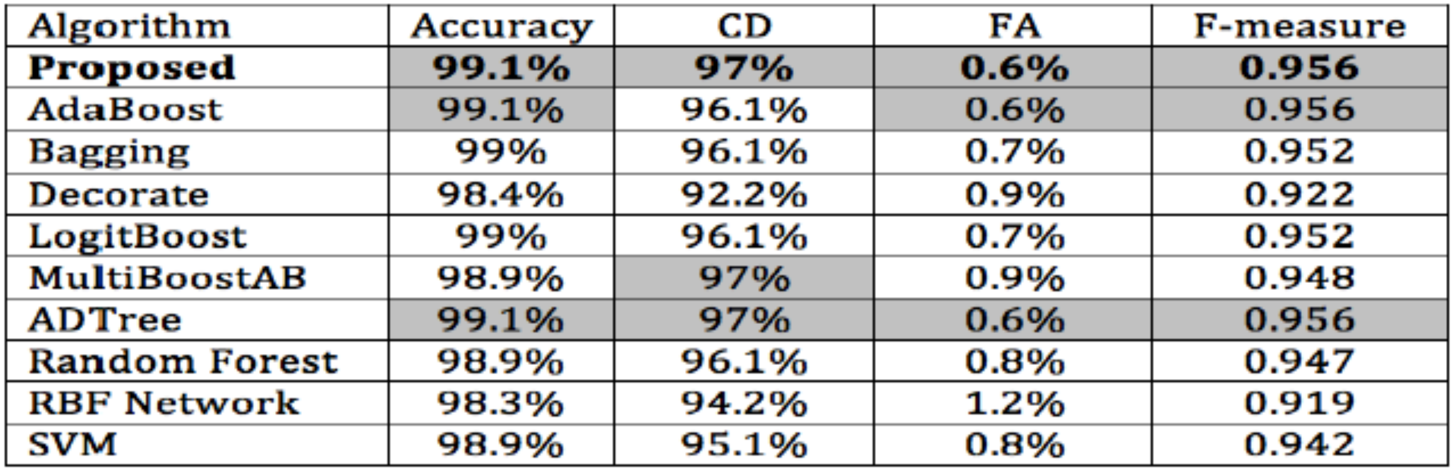
(a)



Input features	Accuracy	CD	FA	F-measure
# of answers and # of best answers	97.6%	100%	2.6%	0.890
# of best answers and Z-score	92.7%	94.8%	7.5%	0.718
# of best answers and # of votes received	93.6%	92.8%	6.3%	0.739
# of answers and Z-score	89.1%	96.9%	11.7%	0.635
# of answers, # of best answers and Z-score	98.4%	94.8%	1.2%	0.920
# of answers, # of best answers and # of votes received	98.9%	93.9%	0.5%	0.943
# of best answers, # of votes received and Z-score	96.6%	92.8%	2.9%	0.842
All features	**99.2%**	**95.9%**	**0.4%**	**0.959**

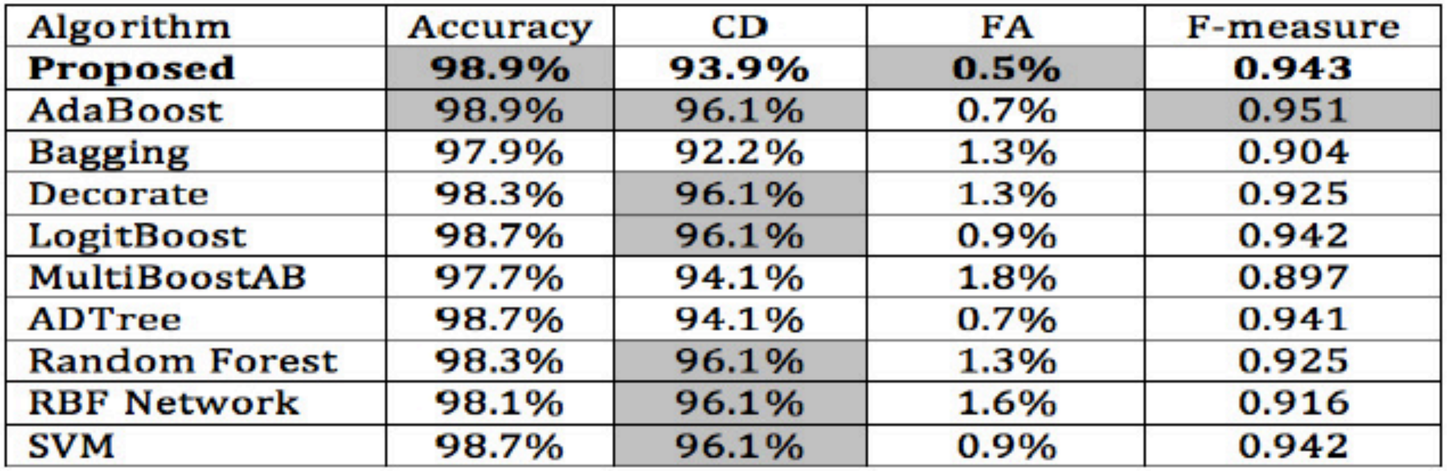
(b)

Performance results of the proposed approach over : (a) Game Development data, (b) Unix & Linux data.



Algorithm	Accuracy	CD	FA	F-measure
Proposed	**99.1%**	**97%**	**0.6%**	**0.956**
AdaBoost	99.1%	96.1%	0.6%	0.956
Bagging	99%	96.1%	0.7%	0.952
Decorate	98.4%	92.2%	0.9%	0.922
LogitBoost	99%	96.1%	0.7%	0.952
MultiBoostAB	98.9%	97%	0.9%	0.948
ADTree	99.1%	97%	0.6%	0.956
Random Forest	98.9%	96.1%	0.8%	0.947
RBF Network	98.3%	94.2%	1.2%	0.919
SVM	98.9%	95.1%	0.8%	0.942

(a)



Algorithm	Accuracy	CD	FA	F-measure
Proposed	**98.9%**	**93.9%**	**0.5%**	**0.943**
AdaBoost	98.9%	96.1%	0.7%	0.951
Bagging	97.9%	92.2%	1.3%	0.904
Decorate	98.3%	96.1%	1.3%	0.925
LogitBoost	98.7%	96.1%	0.9%	0.942
MultiBoostAB	97.7%	94.1%	1.8%	0.897
ADTree	98.7%	94.1%	0.7%	0.941
Random Forest	98.3%	96.1%	1.3%	0.925
RBF Network	98.1%	96.1%	1.6%	0.916
SVM	98.7%	96.1%	0.9%	0.942

(b)

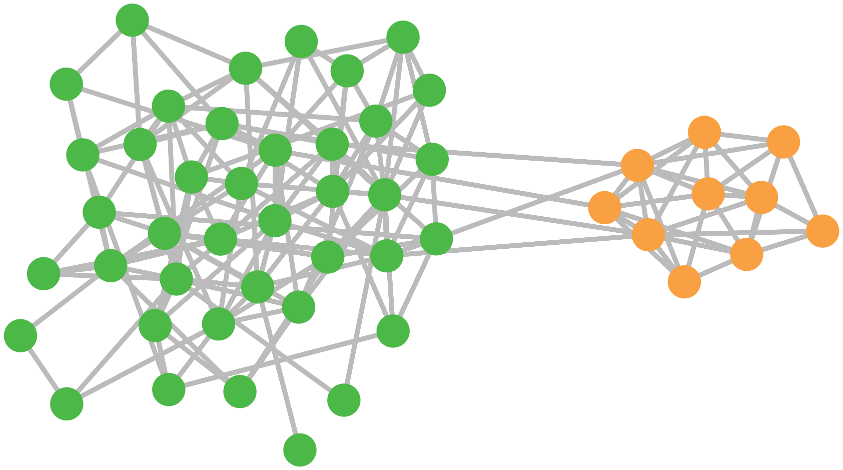
Accuracies of compared algorithms on: (a) Game Development data, (b) Unix & Linux data.

SNA Techniques

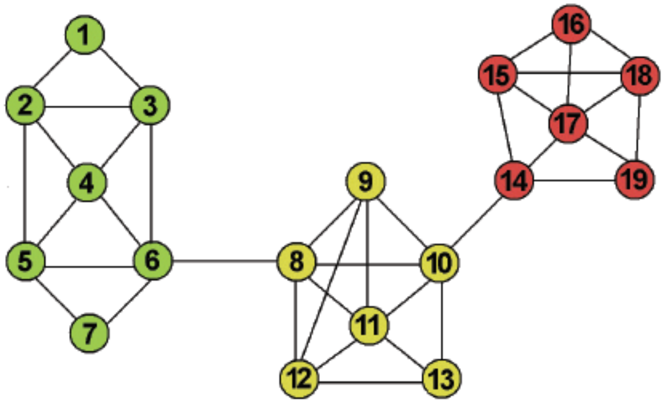
Prominent problems

- Social network extraction/construction
- Identifying prominent/trusted/expert actors
- Identifying Spammers
- **Discovering communities in social networks**
- Link prediction
- Approximating large social networks
- Evolution of social networks

Community Structure in Social Network



Graph Clustering



0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	0	1	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	0	1	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0

Algorithms based on Czekanovski-Dice Distance

Distance between two nodes

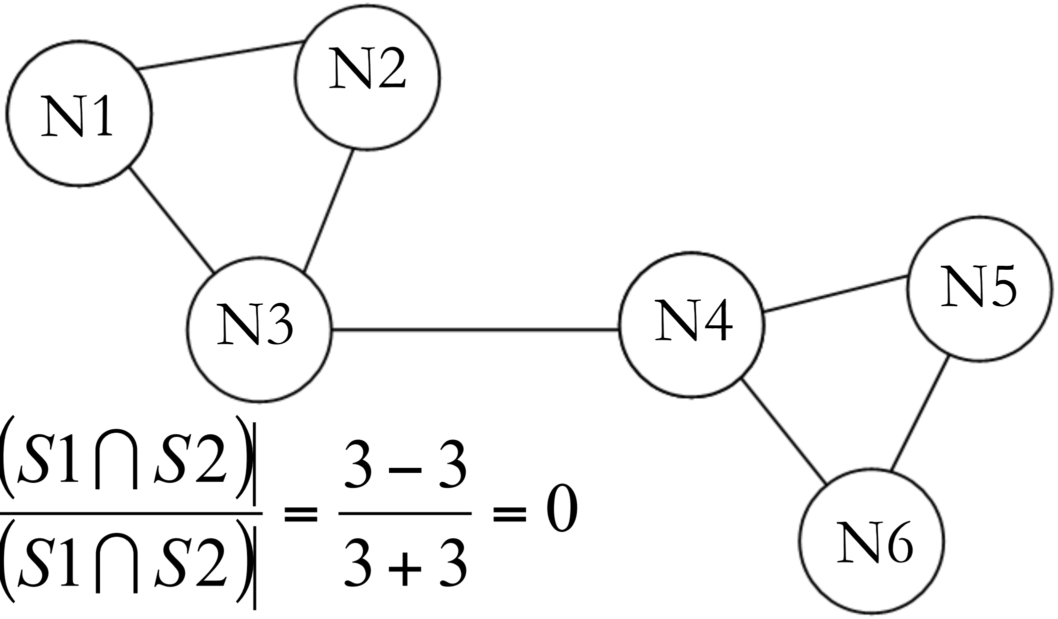
$$dist(N1, N2) = \frac{\left|(S1 \cup S2)\right| - \left|(S1 \cap S2)\right|}{\left|(S1 \cup S2)\right| + \left|(S1 \cap S2)\right|}$$

S1: number of nodes connected to N1 (including N1)
S2: number of nodes connected to N2 (including N2)

Small distance → High similarity

Czekanovski-Dice Distance

- **Exemple**



- dist(N1, N2) = ?

S1 = {N1, N2, N3}
S2 = {N2, N1, N3}

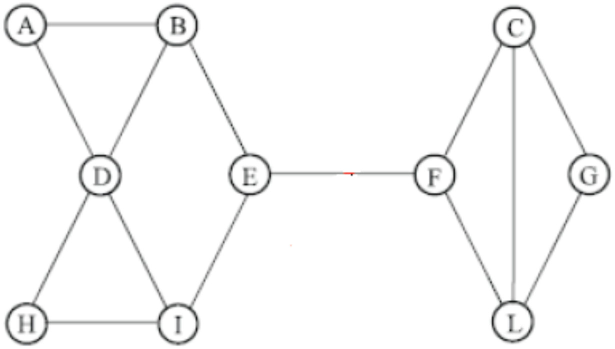
$$dist(N1,N2)=\frac{|(S1\cup S2)|-|(S1\cap S2)|}{|(S1\cup S2)|+|(S1\cap S2)|}=\frac{3-3}{3+3}=0$$

- dist(N3, N4) = ?

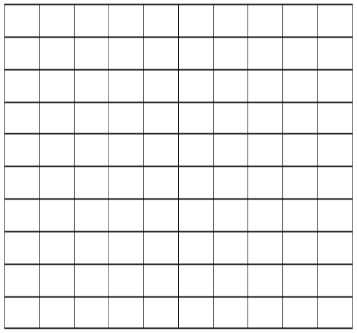
S3 = {N3, N1, N2, N4}
S4 = {N4, N3, N5, N6}

$$dist(N3,N4)=\frac{|(S3\cup S4)|-|(S3\cap S4)|}{|(S3\cup S4)|+|(S3\cap S4)|}=\frac{6-2}{6+2}=0.5$$

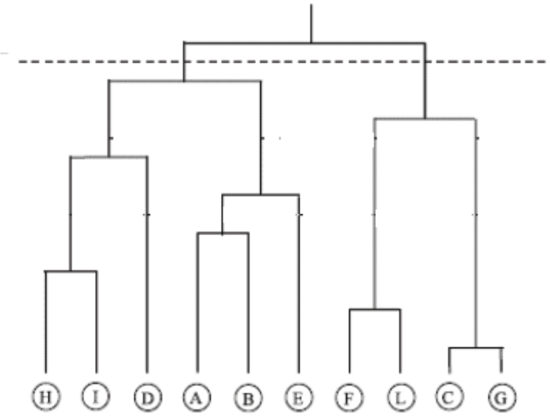
Czekanovski-Dice Distance



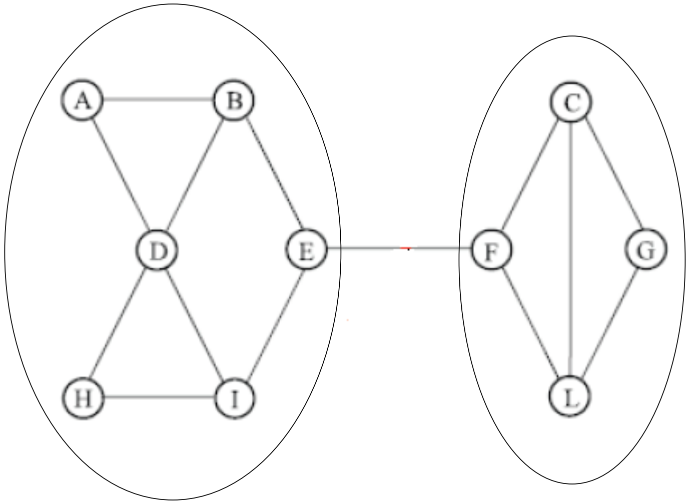
(a) Graph



(b) Smilarity Matrix



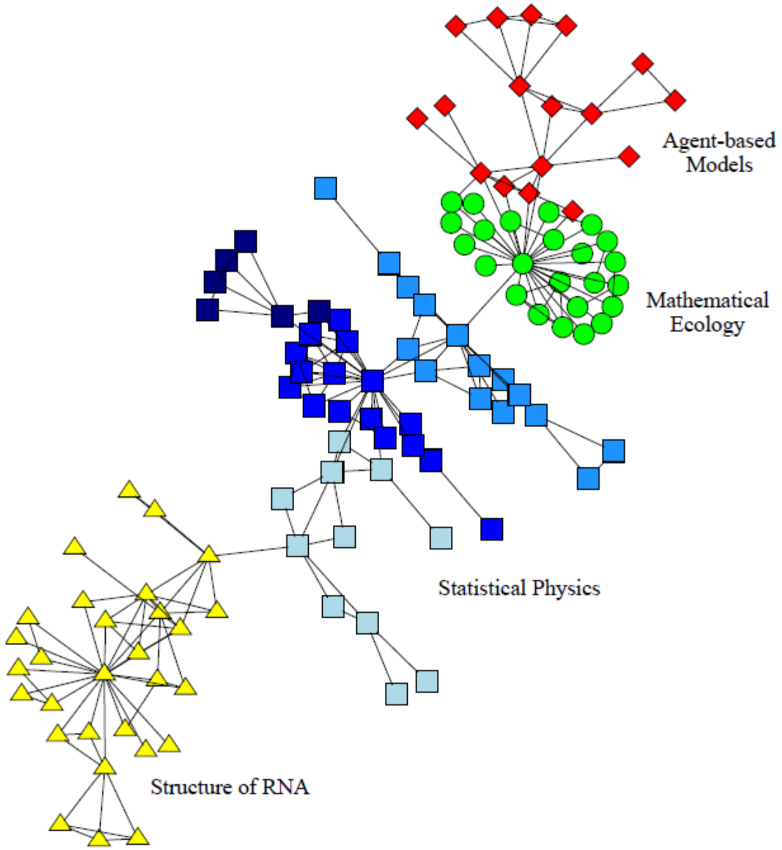
(c) Dendogramme



(d) Clustering

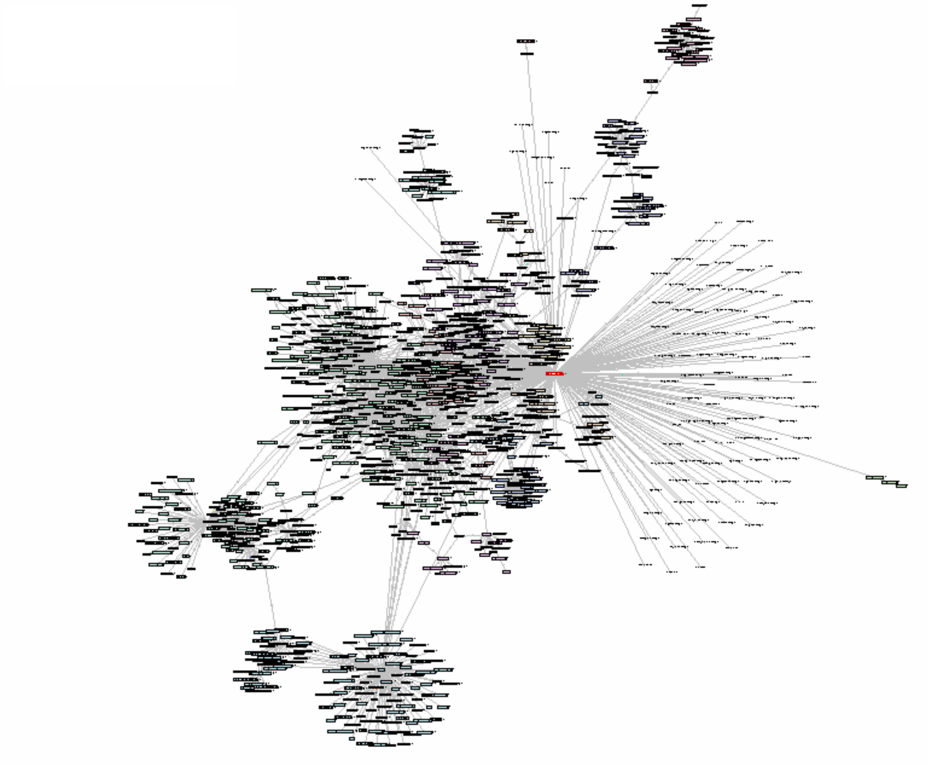
Application

The Santa Fe Institute collaboration network



Application

Enron email network



Discovering Knowledge-Sharing Communities in Question-Answering Forums

Knowledge-Sharing Community

1. A knowledge-sharing community is defined by a set of askers and authoritative users.
2. Within each community, askers exhibit more homogenous behavior in terms of their interactions with authoritative users than elsewhere.
3. Authoritative users may belong to more than one community.

Knowledge-Sharing Community

Existing graph-based community detection methods are not appropriate for our study.

Example

$a_1 : e_1, e_2$

$a_2 : e_1, e_2$

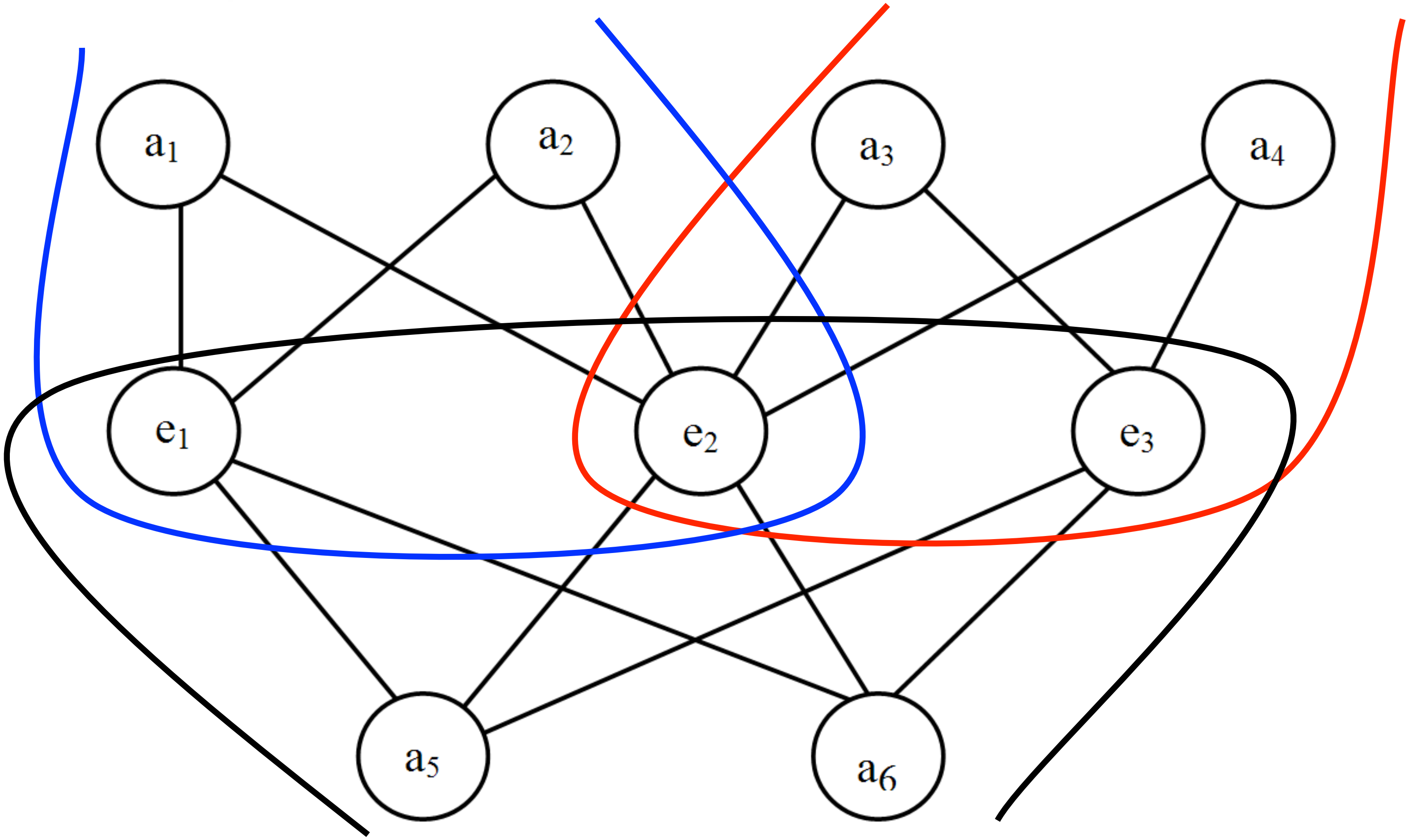
$a_3 : e_2, e_3$

$a_4 : e_2, e_3$

$a_5 : e_1, e_2, e_3$

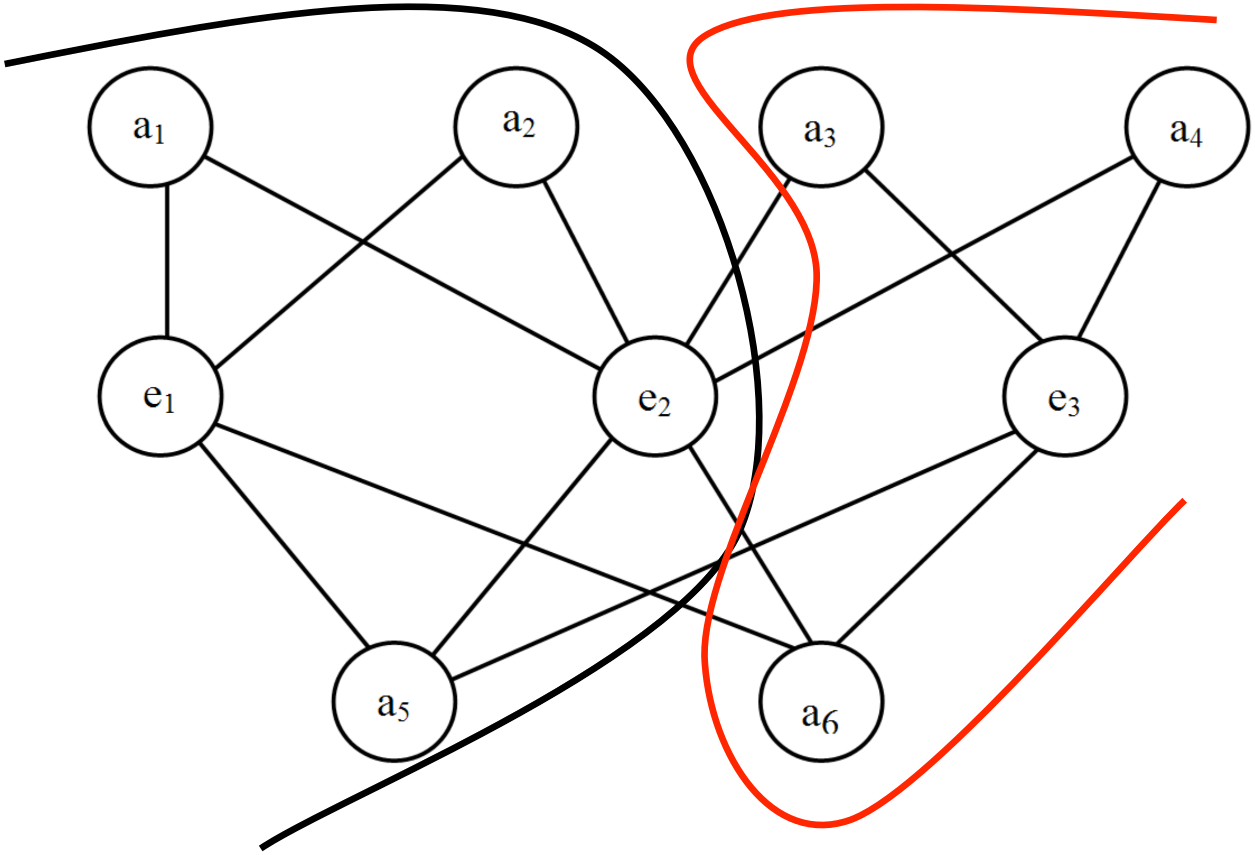
$a_6 : e_1, e_2, e_3$

Example



Modeling users interactions as a graph

The GRACLUS Algorithm



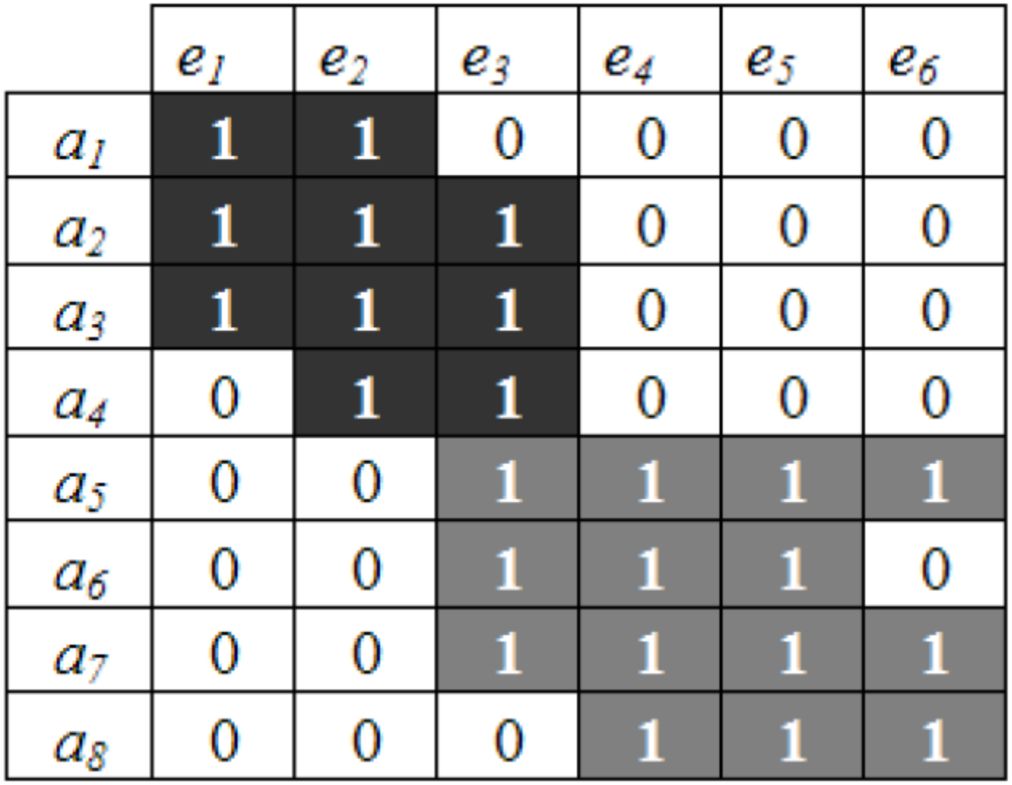
Modeling Interactions Between Users

➢ We use a transactional data model to represent the interactions between askers and authoritative users.

$$
\begin{aligned}
T_1 &= \{e_1, e_2\} \\
T_2 &= \{e_1, e_2, e_3\} \\
T_3 &= \{e_1, e_2, e_3\} \\
T_4 &= \{e_2, e_3\}
\end{aligned}
\qquad
\begin{aligned}
T_5 &= \{e_3, e_4, e_5, e_6\} \\
T_6 &= \{e_3, e_4, e_5\} \\
T_7 &= \{e_3, e_4, e_5, e_6\} \\
T_8 &= \{e_4, e_5, e_6\}
\end{aligned}
$$

- The first community is defined by *T1, T2, T3* et *T4*
- The second community is defined by *T5, T6, T7* et *T8*

Illustration



	e_1	e_2	e_3	e_4	e_5	e_6
a_1	1	1	0	0	0	0
a_2	1	1	1	0	0	0
a_3	1	1	1	0	0	0
a_4	0	1	1	0	0	0
a_5	0	0	1	1	1	1
a_6	0	0	1	1	1	0
a_7	0	0	1	1	1	1
a_8	0	0	0	1	1	1

Boolean representation of the interaction between askers and authoritative users.

The TRANCLUS Algorithm

- $A = \{a_1, a_2, \ldots, a_n\}$ a set of n askers
- $E = \{e_1, e_2, \ldots, e_d\}$ a set of d authoritative users
- $TD = \{T_1, T_2, \ldots, T_n\}$ a collection of n transactions that summarizes the interactions of all askers a_i with the identified authoritative users.

Problem Definition

Given the set *A* of askers and the set *E* of authoritative users,

- Construct the set *TD*.

- Partition *TD* into a set of disjoint clusters

  $C = \{C_1, C_2, \ldots, C_{nc}\}$

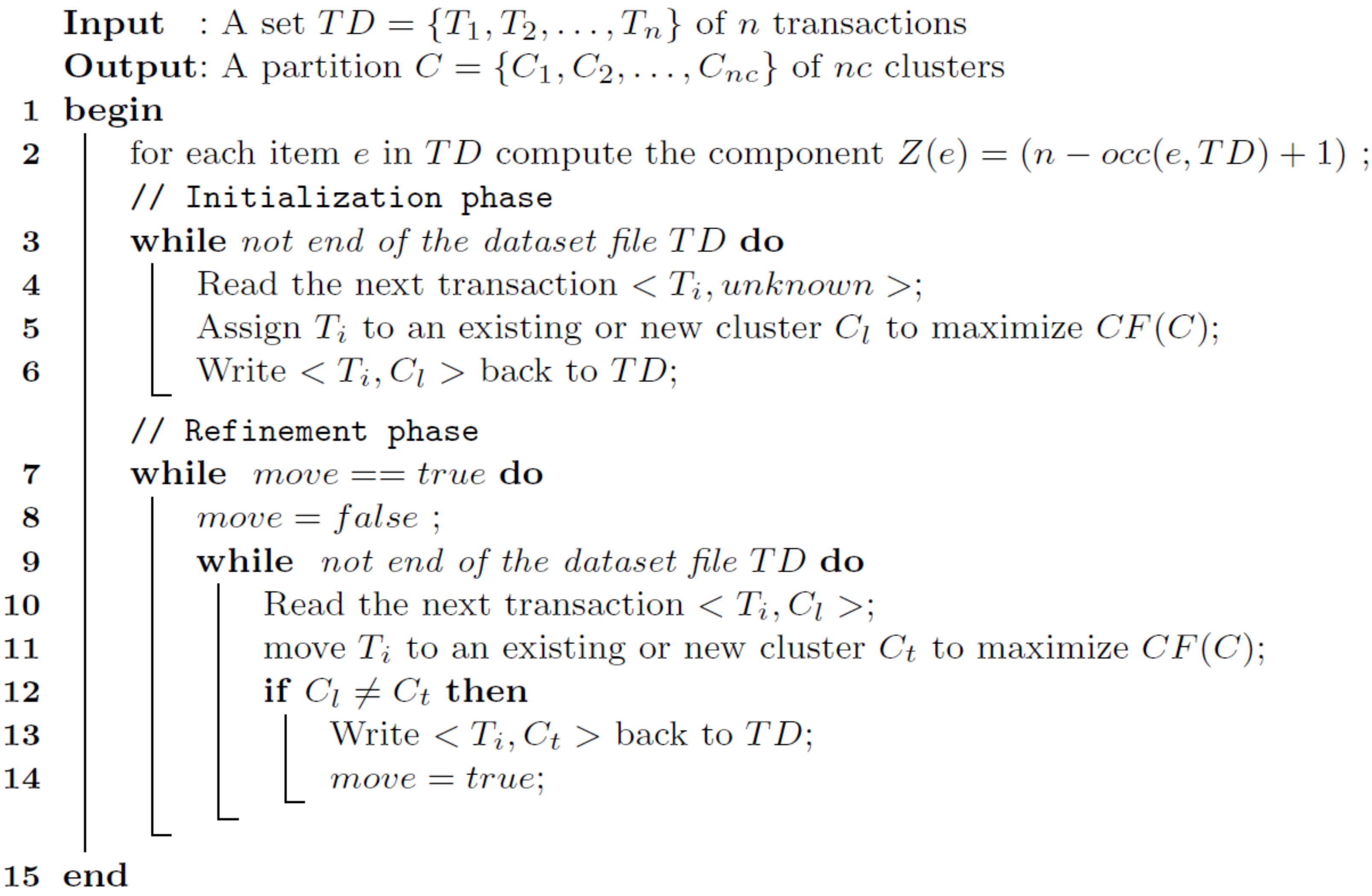
➢ The identified clusters represent the communities we want to discover.

Criterion Function

$$CF(C) = \frac{1}{n^2} \sum_{s=1}^{nc} \left[\frac{1}{n_s} \sum_{e \in C_s} \left(\left(occ(e, C_s) \right)^3 \times Z(e) \right) \right]$$

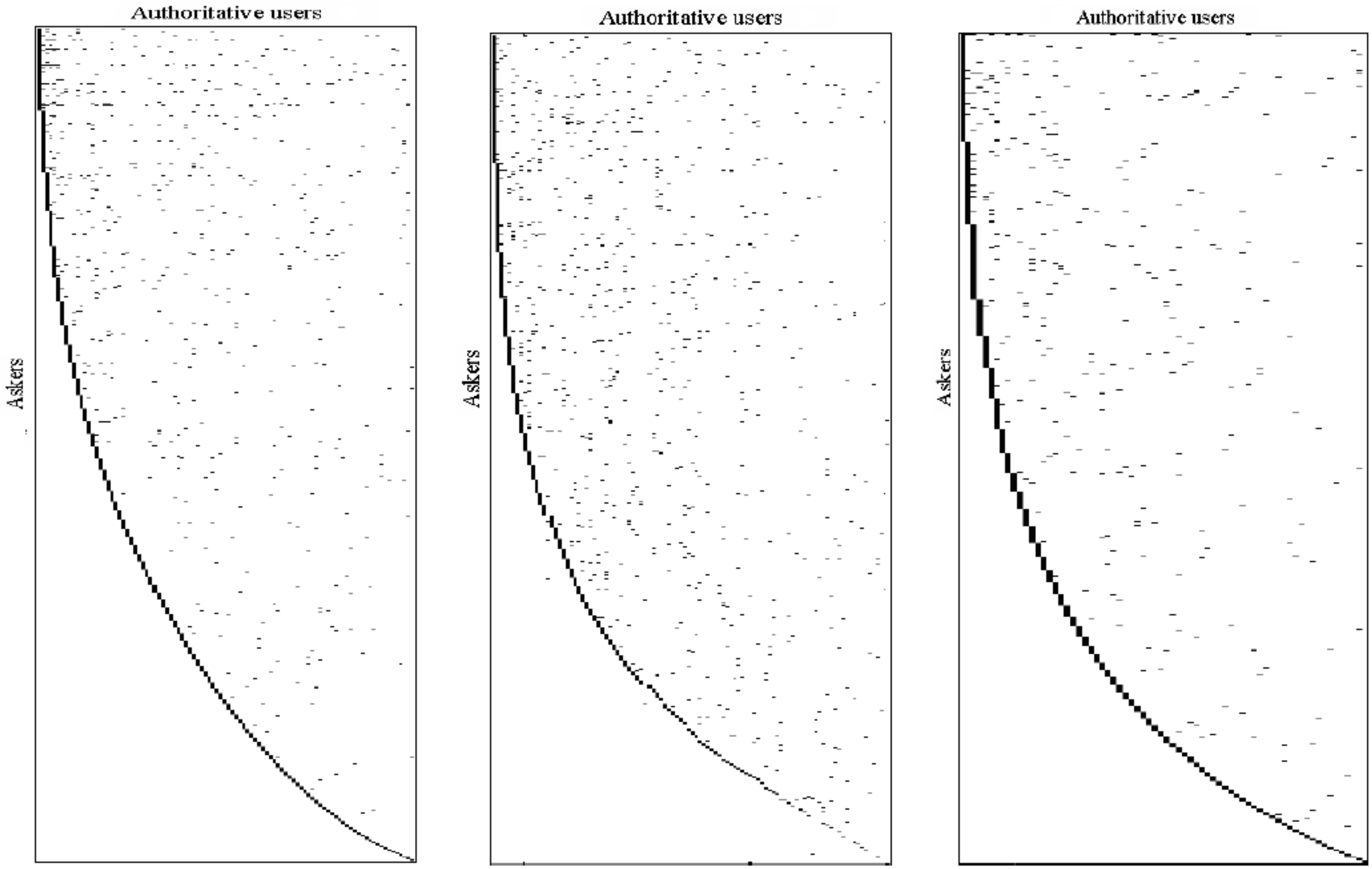
$$Z(e) = \left(n - occ(e, TD) + 1 \right)$$

The TRANCLUS Scheme



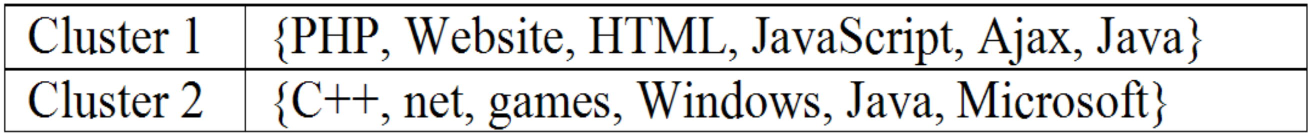
```
Input  : A set $TD = \{T_1, T_2, \ldots, T_n\}$ of $n$ transactions
Output: A partition $C = \{C_1, C_2, \ldots, C_{nc}\}$ of $nc$ clusters
begin
    for each item $e$ in $TD$ compute the component $Z(e) = (n - occ(e, TD) + 1)$ ;
    // Initialization phase
    while not end of the dataset file $TD$ do
        Read the next transaction $< T_i, unknown >$;
        Assign $T_i$ to an existing or new cluster $C_l$ to maximize $CF(C)$;
        Write $< T_i, C_l >$ back to $TD$;
    // Refinement phase
    while $move == true$ do
        $move = false$ ;
        while not end of the dataset file $TD$ do
            Read the next transaction $< T_i, C_l >$;
            move $T_i$ to an existing or new cluster $C_t$ to maximize $CF(C)$;
            if $C_l \neq C_t$ then
                Write $< T_i, C_t >$ back to $TD$;
                $move = true$;
end
```

Application to Yahoo! Answers



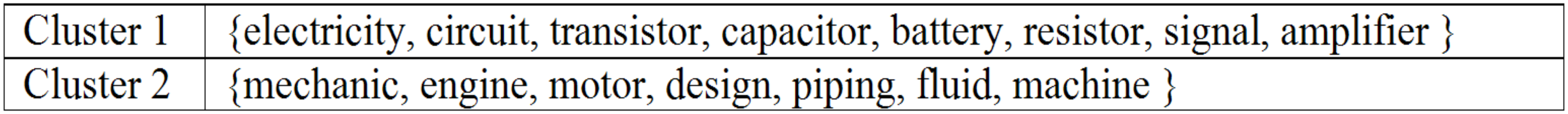
(a) Biology. (b) Chemistry. (c) Engineering.

Content Analysis



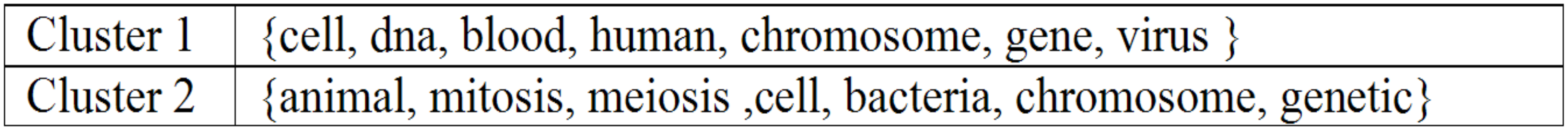
Cluster 1	{PHP, Website, HTML, JavaScript, Ajax, Java}
Cluster 2	{C++, net, games, Windows, Java, Microsoft}

(a) Programming.



Cluster 1	{electricity, circuit, transistor, capacitor, battery, resistor, signal, amplifier }
Cluster 2	{mechanic, engine, motor, design, piping, fluid, machine }

(b) Engineering.



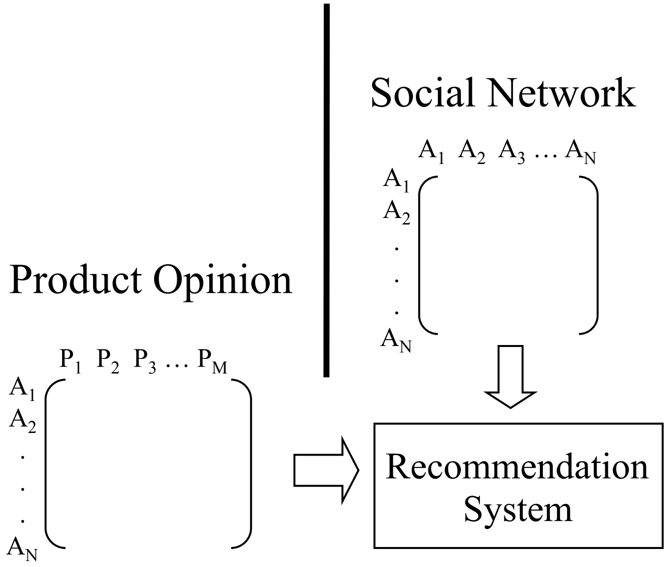
Cluster 1	{cell, dna, blood, human, chromosome, gene, virus }
Cluster 2	{animal, mitosis, meiosis ,cell, bacteria, chromosome, genetic}

(c) Biology.

➢ The clustered askers tend to post questions on closed related topics

Emerging Application

Influence of Social Networks on Product Recommendations



Social Network

Product Opinion

Recommendation System

- Understanding the impact of social networks on market behavior
- Improved recommendation systems