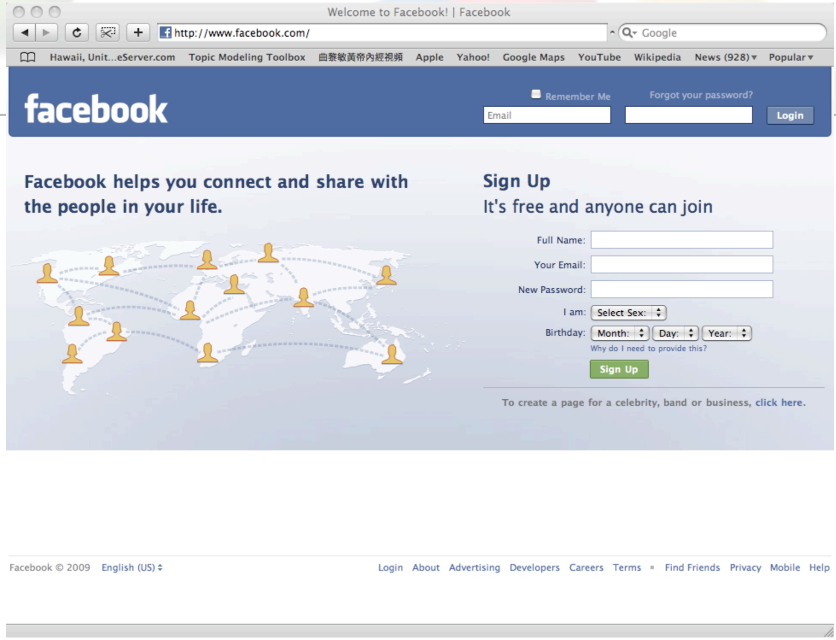
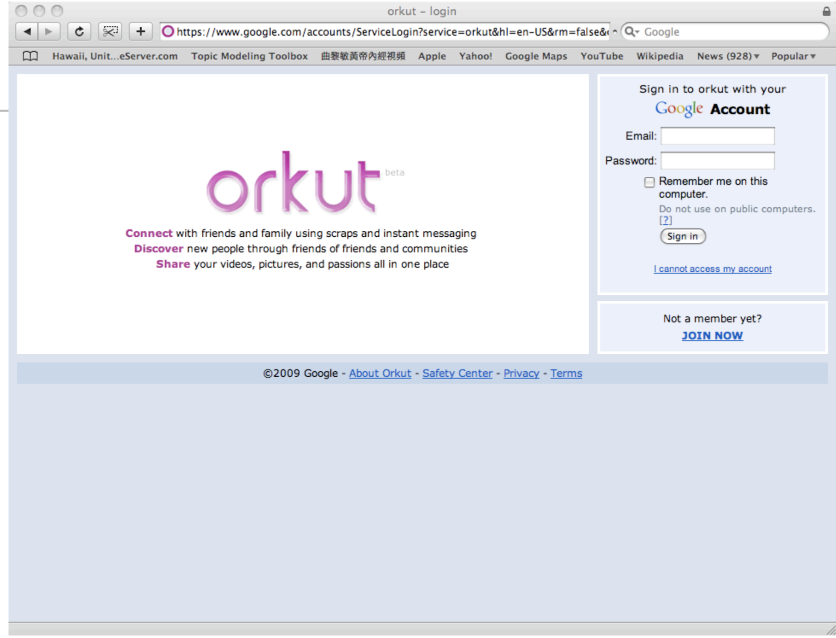
Collaborative Filtering for Orkut Communities: Discovery of User Latent Behavior

Wen-Yen Chen
Computer Science
University of California, Santa Barbara

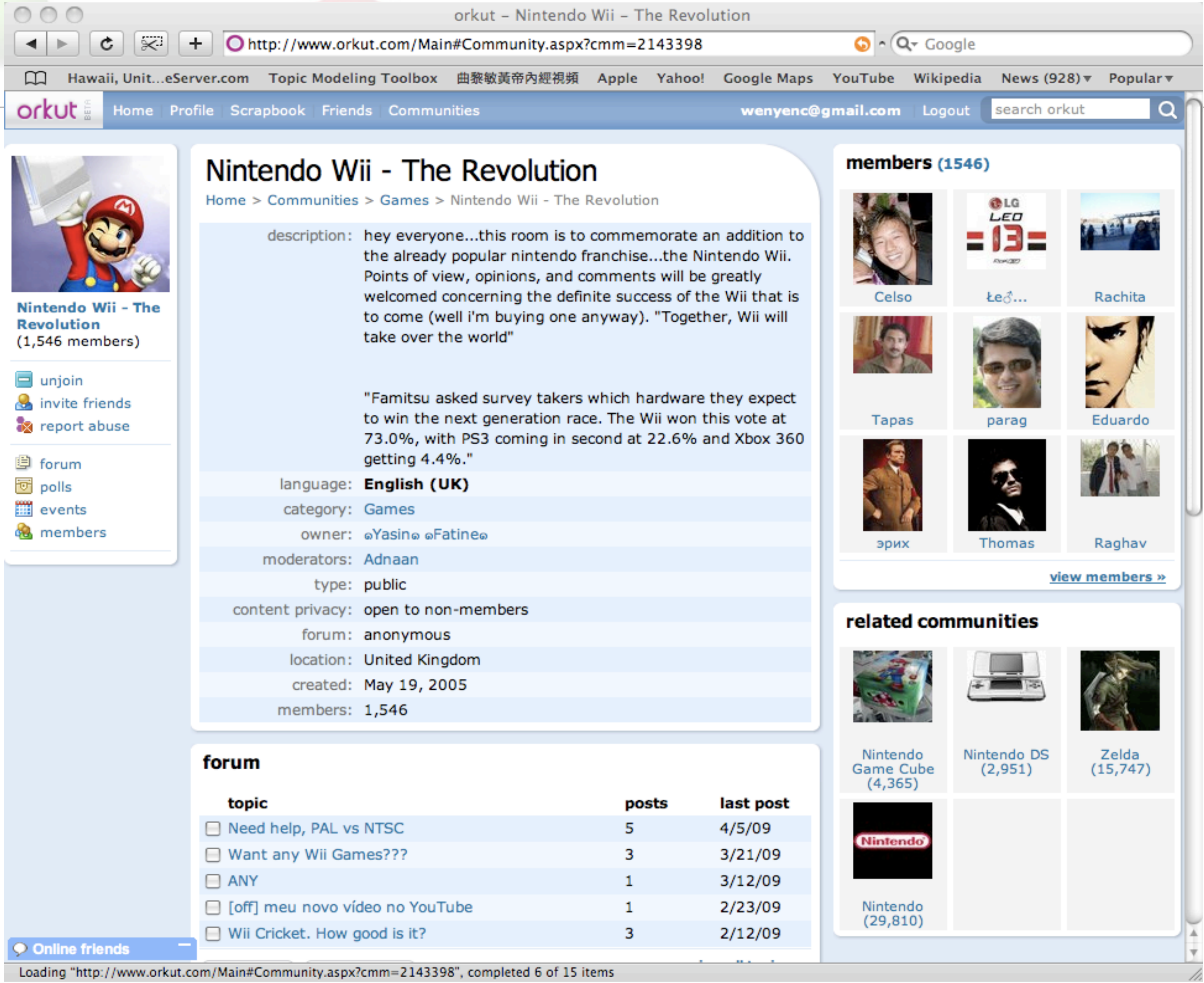
Joint work with
Jon Chu (MIT)
Junyi Luan (PKU)
Hongjie Bai (Google)
Edward Chang (Google)


Welcome to Facebook! | Facebook
http://www.facebook.com/
facebook
Remember Me
Forgot your password?
Login
Facebook helps you connect and share with the people in your life.
Sign Up
It's free and anyone can join
Full Name:
Your Email:
New Password:
I am: Select Sex:
Birthday: Month: Day: Year:
Why do I need to provide this?
Sign Up
To create a page for a celebrity, band or business, click here.
Facebook © 2009 English (US)
Login About Advertising Developers Careers Terms Find Friends Privacy Mobile Help


orkut - login
https://www.google.com/accounts/ServiceLogin?service=orkut&hl=en-US&rm=false&
orkut
Connect with friends and family using scraps and instant messaging
Discover new people through friends of friends and communities
Share your videos, pictures, and passions all in one place
Sign in to orkut with your
Google Account
Email:
Password:
Remember me on this computer.
Do not use on public computers.
Sign in
I cannot access my account
Not a member yet?
JOIN NOW
©2009 Google - About Orkut - Safety Center - Privacy - Terms


MySpace
http://www.myspace.com/
myspace.com
¿Eres Nuevo en MySpace?
Date una vuelta
Buscar
Inicio Explorar Buscar gente Foros Música Video Más
Registrarse
FILMO, LUEGO EXISTO.
Entra y participa
MOVIL FILM FEST
SI FILMAS PARTICIPA EN EL MOVIL FILM FEST
CATEGORIA MUSICAL Y
GANA EL PREMIO MYSPACE
New!
¡pincha aquí!
O ENTRA EN
www.myspace.com/amigosdefangoria
Hoy en MySpace
Noticias Música Video destacado Gente
¡Consigue tu "Selector Pass": Entrada + Sorpresas!
Participa y gana tu "Selector Pass" para los próximos conciertos de Lagartija Nick.
Iniciar sesión
Regístrate!
Correo:
Contraseña:
Recordar mis datos
Iniciar sesión
¿Has olvidado tu contraseña?


Twitter: What are you doing?
http://twitter.com/
twitter
Select Language ...
What is Twitter?
What? Why? How?
Watch a video!
Please sign in
user name or email address:
password:
Remember me
Sign In
Forgot password? Click here.
Already using Twitter from your phone? Click here.
Twitter is a service for friends, family, and co-workers to communicate and stay connected through the exchange of quick, frequent answers to one simple question: What are you doing?
Get Started—Join!
When I first started doing it, I thought, 'geez, not another website to worry about updating and checking', but now I'm glad I did it.
point_chevalier, livejournaller
If you aren't familiar with Twitter, it is one of those things, like MySpace, that sounds totally ridiculous and stupid when you first hear about it. But once you start using it, you realize how much fun it is.
It's one of the fastest-growing phenomena on the Internet.
New York Times


orkut – Nintendo Wii – The Revolution
http://www.orkut.com/Main#Community.aspx?cmm=2143398
Google
Hawaii, Unit...eServer.com
Topic Modeling Toolbox
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News (928)
Popular
orkut
Home
Profile
Scrapbook
Friends
Communities
wenyenc@gmail.com
Logout
search orkut
Nintendo Wii - The Revolution
Home > Communities > Games > Nintendo Wii - The Revolution
description: hey everyone...this room is to commemorate an addition to the already popular nintendo franchise...the Nintendo Wii. Points of view, opinions, and comments will be greatly welcomed concerning the definite success of the Wii that is to come (well i'm buying one anyway). "Together, Wii will take over the world"
"Famitsu asked survey takers which hardware they expect to win the next generation race. The Wii won this vote at 73.0%, with PS3 coming in second at 22.6% and Xbox 360 getting 4.4%."
language: English (UK)
category: Games
owner: ๑Yasin๑ ๑Fatine๑
moderators: Adnaan
type: public
content privacy: open to non-members
forum: anonymous
location: United Kingdom
created: May 19, 2005
members: 1,546
Nintendo Wii - The Revolution
(1,546 members)
unjoin
invite friends
report abuse
forum
polls
events
members
members (1546)
Celso
Rachita
Tapas
parag
Eduardo
эрих
Thomas
Raghav
view members »
related communities
Nintendo Game Cube (4,365)
Nintendo DS (2,951)
Zelda (15,747)
Nintendo (29,810)
forum
topic
posts
last post
Need help, PAL vs NTSC
5
4/5/09
Want any Wii Games???
3
3/21/09
ANY
1
3/12/09
[off] meu novo vídeo no YouTube
1
2/23/09
Wii Cricket. How good is it?
3
2/12/09
Online friends
Loading "http://www.orkut.com/Main#Community.aspx?cmm=2143398", completed 6 of 15 items

Motivation

Social-network sites are popular and attract millions of users a day

- Facebook, Orkut, Myspace,Twitter…
- Orkut has more than 130M users, 30M communities, 10K communities created daily

Rapid growth of user-generated data available

- Communities, images, videos, posts, friendships…
- Information overload problem

We focus on personalized community recommendation task

- Collaborative filtering (CF) approach

Collaborative Filtering (CF)

The operative assumption underlying collaborative filtering

- Users who were similar in the past are likely to be similar in the future
- Use similar users' behaviors to make recommendations

Algorithms of three different types

- Memory-based
- Model-based
- Association rules

Collaborative Filtering for Orkut Communities

Investigate two algorithms from very different domains

- Association rules mining (ARM)
  - Discover associations between communities (explicit relations)
  - Users joining “NYY” usually join “MLB”, rule: NYY → MLB
  - Target user joins “NYY”, being recommended “MLB”
  - Fewer common users between “New York Mets” and “MLB”, no rules

New York Yankees ***Major League Baseball*** ***New York Mets***

Collaborative Filtering for Orkut Communities

Investigate two algorithms from very different domains

- Association rules mining (ARM)
  - Discover associations between communities (explicit relations)
  - Users joining “NYY” usually join “MLB”, rule: NYY → MLB
  - Target user joins “NYY”, being recommended “MLB”
  - Fewer common users between “New York Mets” and “MLB”, no rules
- Latent Dirichlet Allocation (LDA)
  - Model user-community using latent aspects (implicit relations)
  - Implicit relation exists between “NYM” and “MLB” via latent structure

Formulate ARM to Community Recommendation

View user as a transaction and his joined communities as items

User	Communities
u_1	$\{c_1, c_3, c_7\}$
u_2	$\{c_3, c_7, c_8, c_9\}$
u_3	$\{c_2, c_3, c_8\}$
u_4	$\{c_1, c_8, c_9\}$

FP-growth

$Supp_{threshold} = 2$

Frequent Itemsets	Support
$\{c_1\}$	2
$\{c_3\}$	3
$\{c_7\}$	2
$\{c_8\}$	3
$\{c_9\}$	2
$\{c_3, c_7\}$	2
$\{c_3, c_8\}$	2
$\{c_8, c_9\}$	2

Association Rules	Support	Confidence
$c_3 \Rightarrow c_7$	2	66.7%
$c_3 \Rightarrow c_8$	2	66.7%
$c_7 \Rightarrow c_3$	2	100%
$c_8 \Rightarrow c_3$	2	66.7%
$c_8 \Rightarrow c_9$	2	66.7%
$c_9 \Rightarrow c_8$	2	100%

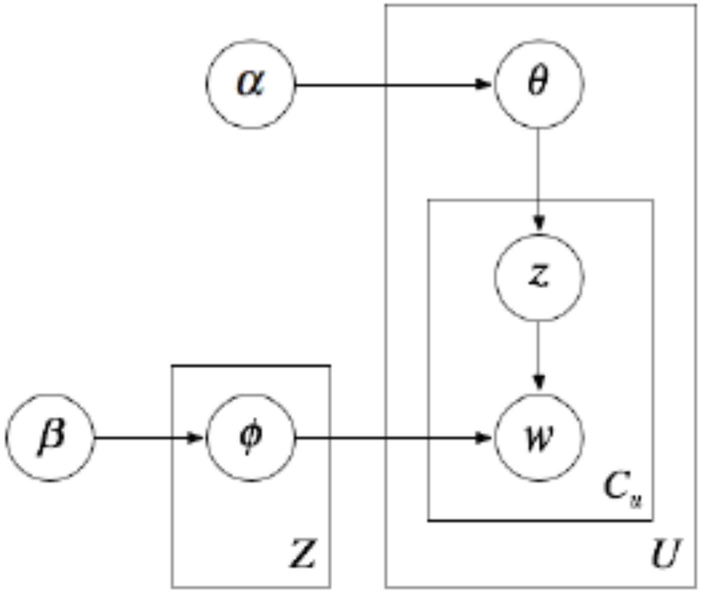
- supp(A) = # of transactions containing A
- supp(A=>B) = supp(A,B)
- conf(A=>B) = supp(A,B) / supp(A)

Recommendation based on rules

- If joining (c_7, c_8), being recommended c_3 (1.667) and c_9 (0.667)

Formulate LDA to Community Recommendation

View users as docs, communities as words and membership counts as co-occurrence counts



- α, β: symmetric Dirichlet priors
- θ: per-user topic distribution
- ϕ: per-topic community distribution

Gibbs sampling

$$P(z_i = j | w_i = c, \mathbf{z}_{-i}, \mathbf{w}_{-i}) \propto \underbrace{\frac{C^{CZ}_{cj} + \beta}{\sum_{c'} C^{CZ}_{c'j} + M\beta}}_{\boldsymbol{\phi}} \underbrace{\frac{C^{UZ}_{uj} + \alpha}{\sum_{j'} C^{UZ}_{uj'} + K\alpha}}_{\boldsymbol{\theta}}$$

Recommendations based on learned model parameters

- $\xi_{cu} = \sum_z \phi_{cz} \theta_{zu}$

Parallelization

We parallelized both ARM and LDA

- Parallel ARM effort [RecSys'08]
- Focus more on parallel LDA

We have two parallel frameworks

- MapReduce
- Message Passing Interface (MPI)

MapReduce and MPI

MapReduce

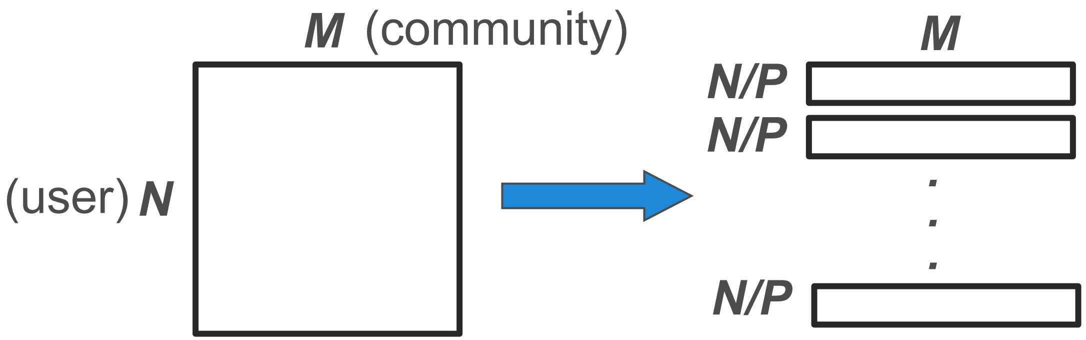
- User specified Map and Reduce functions
- Map: generates a set of intermediate key/value pairs
- Reduce: reduce the intermediate values with the same key
- Read/Write data using disk I/O
- Intensive I/O cost but provide fault-tolerance mechanism

Message Passing Interface (MPI)

- Send/receive data to/from machine's memory
- Machines can communicate via MPI library routines
- Lazy checkpoints for fault-tolerance
- Suitable for algorithms with iterative procedures

Parallelization

We have *P* machines and distribute the computation by rows



Each machine *i*

Community-topic count

User-topic count

- Computes local variables $C_{cj}^{CZ}(i)$ and $C_{uj}^{UZ}(i)$
- Gets global variable $C_{cj}^{CZ} = \sum_i C_{cj}^{CZ}(i)$
  - AllReduce operation

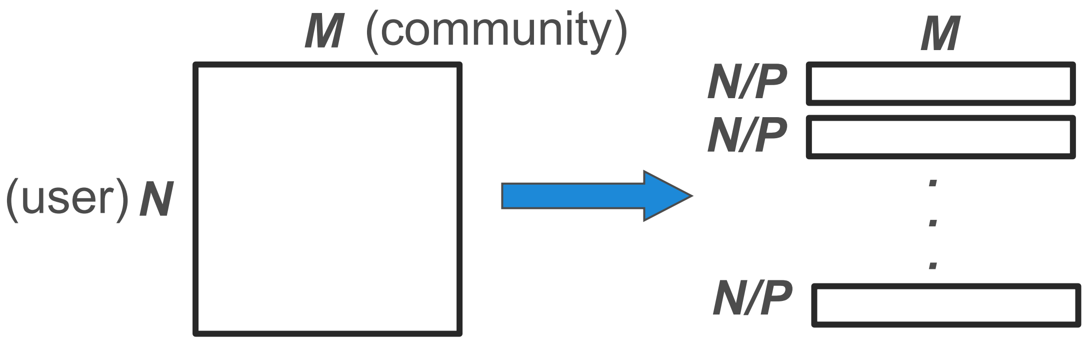
Computation cost

- Before: $O(NLK) \times (\#\text{ of iterations})$
- After: $O(\frac{NLK}{P}) \times (\#\text{ of iterations})$

N: # of users
L: avg # of communities per user
K: # of topics

Parallelization

We have *P* machines and distribute the computation by rows



Each machine *i*

- Computes local variables $C_{cj}^{CZ}(i)$ and $C_{uj}^{UZ}(i)$
- Gets global variable $C_{cj}^{CZ} = \sum_i C_{cj}^{CZ}(i)$

Communication cost

- Communication: $O\left(\alpha \cdot \log P + \beta \cdot \frac{P-1}{P} KM + \gamma \cdot \frac{P-1}{P} KM\right)$

startup time of a transfer *transfer time per unit* *computational time for reduction*

Empirical Study

Orkut data

- Community membership data
- 492,104 users and 118,002 communities
- User/community data are anonymized to preserve privacy

Evaluations

- Recommendation quality using top-*k* ranking metric
- Rank difference between ARM and LDA
- Latent information learned from LDA
- Speedup

Community Recommendation

Evaluation metric

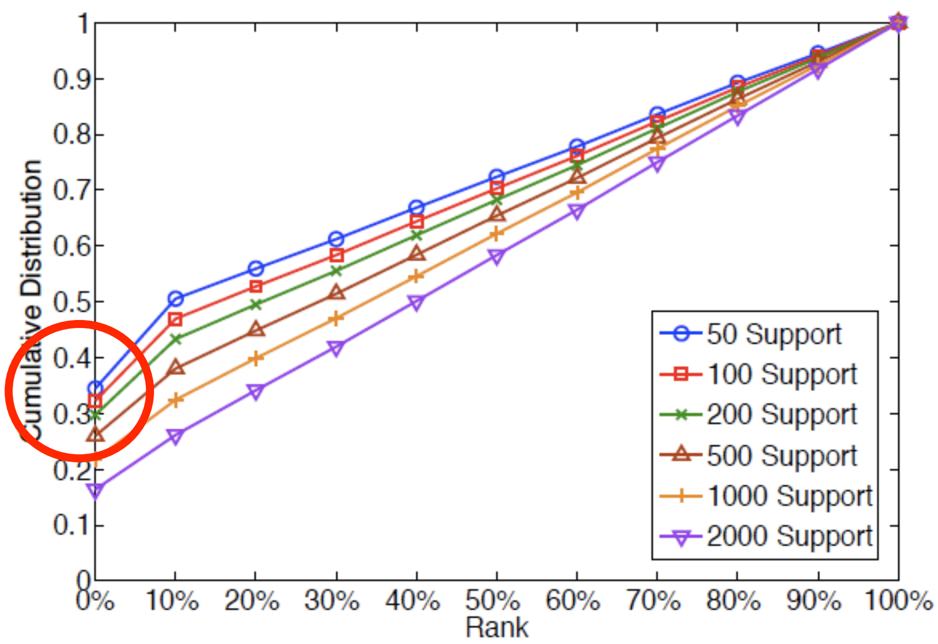
- Output values of two algorithms cannot be compared directly
- Ranking metric: top-*k* recommendation [Y. Koren KDD'08]

Evaluation protocol

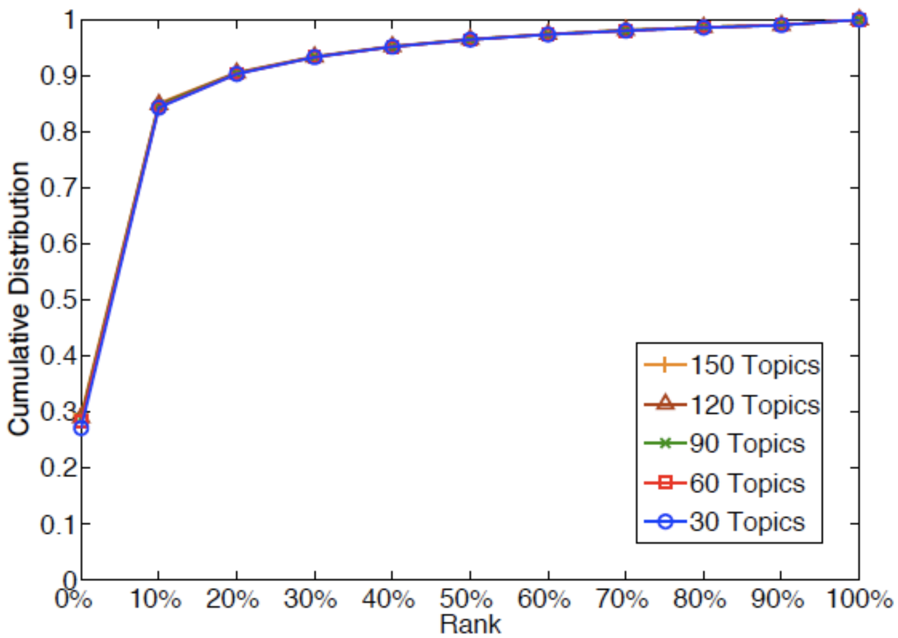
- Randomly withhold one community from user's joined communities
  - Training set for algorithms
- Select k-1 additional random communities not in user's joined communities
- Evaluate set: the withheld community together with k-1 other communities
  - Order the communities by predicted scores
  - Obtain the corresponding rank of the withheld community (0, …, k-1)
- The lower the rank, the more successful the recommendation

Top-*k* recommendation performance

Macro-view (0% - 100%), where k = 1001



ARM



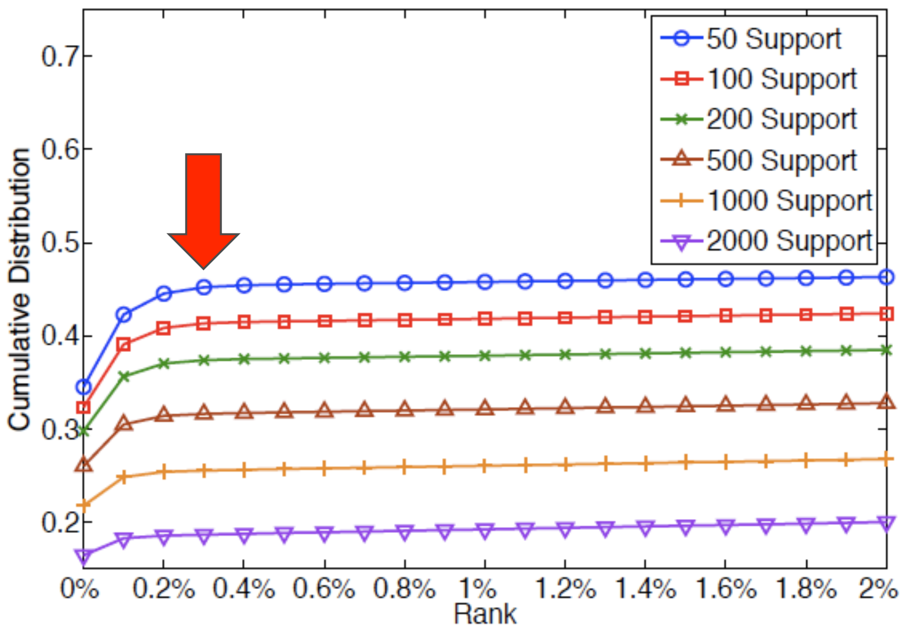
LDA

ARM: higher the support, worse the performance

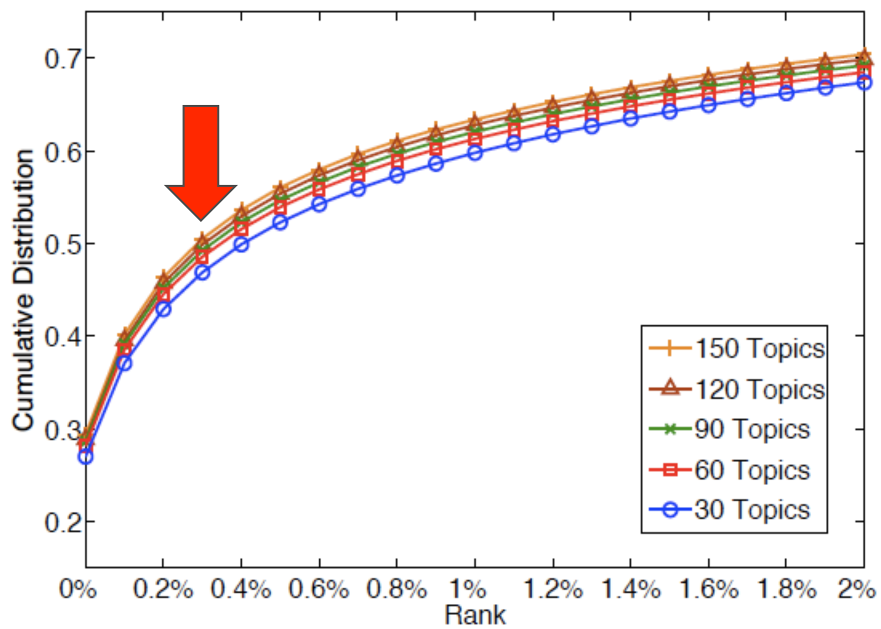
LDA: consistent performance with varying # of topics

Top-*k* recommendation performance (cont.)

Micro-view (0% - 2%), where k = 1001



ARM



LDA

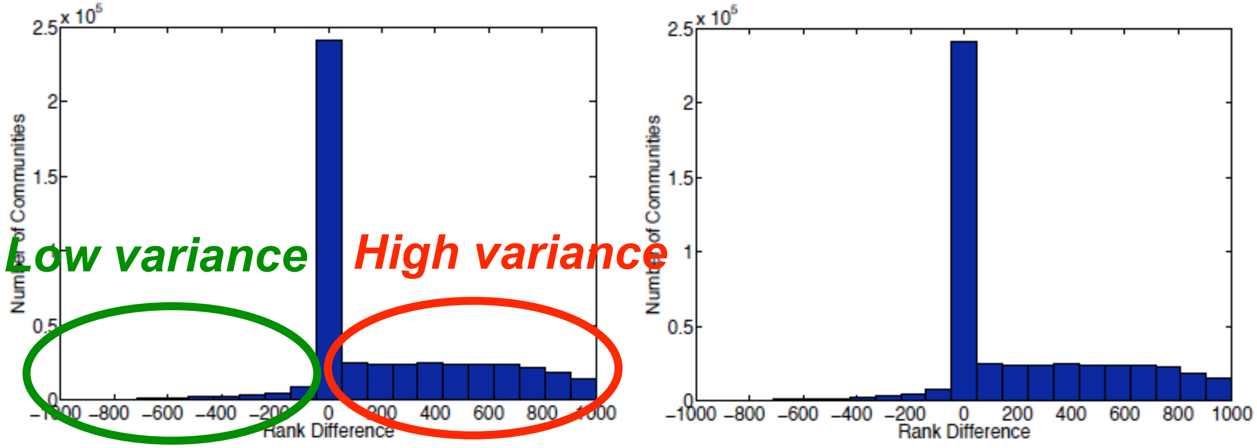
ARM is better when recommending list up to 3 communities

LDA is consistently better when recommending a list of 4 or more

Rank Differences

Rank differences under different parameters

- ARM-50: best-performing ARM
- LDA-30: worst-performing LDA, LDA-150: best-performing LDA
- Rank difference = LHS – RHS



(a) ARM-50 vs. LDA-30 (b) ARM-50 vs. LDA-150

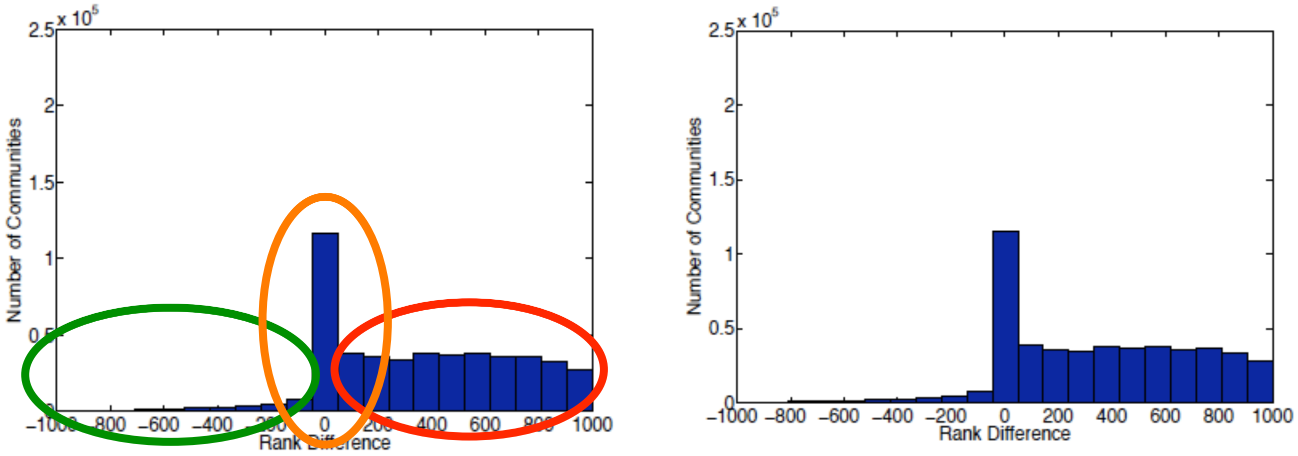
More withhelod communities have positive rank differences

- LDA generally ranks better than ARM
- LDA is better → much better, ARM is better → a little better

Rank Differences (cont.)

Rank differences under different parameters

- ARM-2000: worst-performing ARM
- LDA-30: worst-performing LDA, LDA-150: best-performing LDA



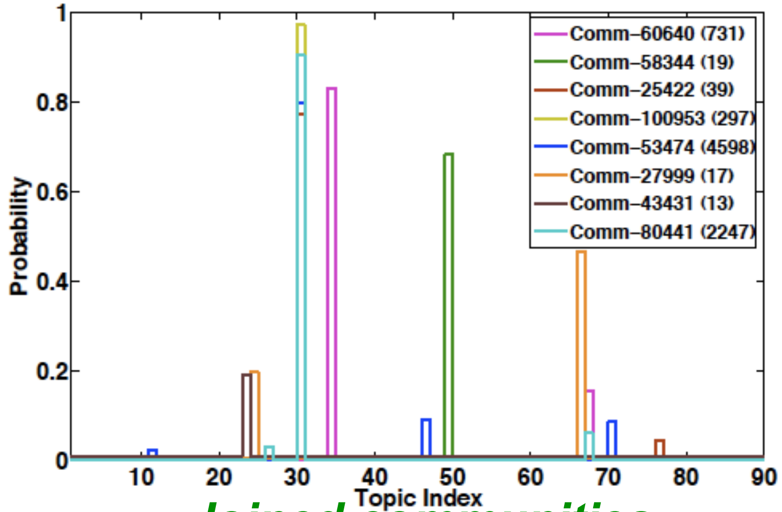
(c) ARM-2000 vs. LDA-30 (d) ARM-2000 vs. LDA-150

Similar patterns but fewer rank difference 0

- Increase in the positive rank difference
- Higher support value causes fewer rules for ARM → narrow coverage

Analysis of Latent Information from LDA (cont.)

User1 whom LDA ranks better

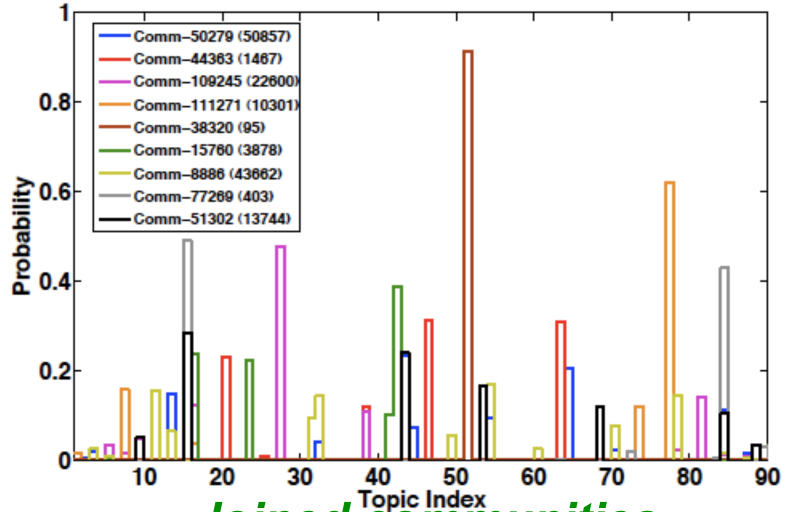


Joined communities

Community #	Community Name	Category	Size
60640	Java certification	Computers/Internet	731
58344	professor Ayaz Isazadeh	Alumni/Schools	19
25422	persiancomputing	Computers/Internet	39
100953	Iranian J2EE developers	Computers/Internet	297
53474	web design	Computers/Internet	4598
27999	Yazd sampad	Schools/Education	17
43431	Tabriz university CS students	Alumni/Schools	13
80441	C#	Computers/Internet	2247
66948	Delphi	Computers/Internet	142

Concentrated topic dist.

User2 whom ARM ranks better



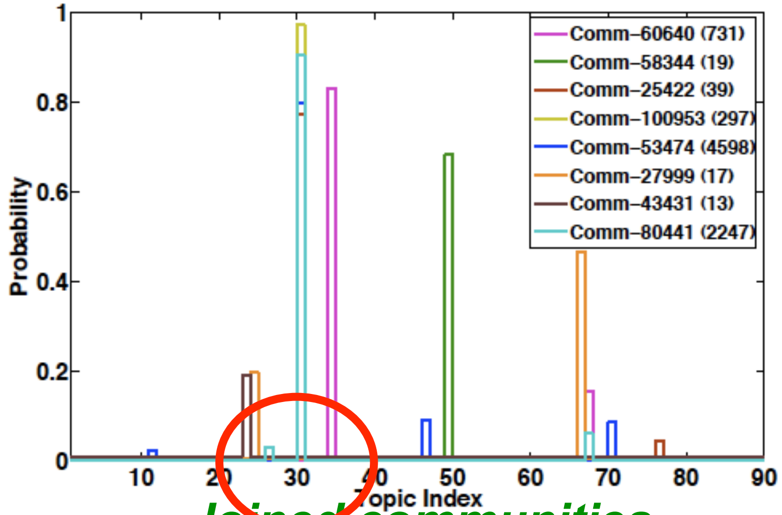
Joined communities

Community #	Community Name	Category	Size
50279	Shahrukh Khan fan club	Individuals	50857
44363	girl power	Religion/Beliefs	1467
109245	love never dies	Romance/Relationships	22600
111271	why friendz break our heart	Romance/Relationships	10301
38320	holy angels school	Alumni/Schools	95
15760	why life is so unpredictable	Other	3878
8886	T20 WC champs	Recreation/Sports	43662
77269	star-one fame serial-remix	Other	403
51302	left right left	Arts/Entertainment	13744
68215	life is too short to live	Other	8197

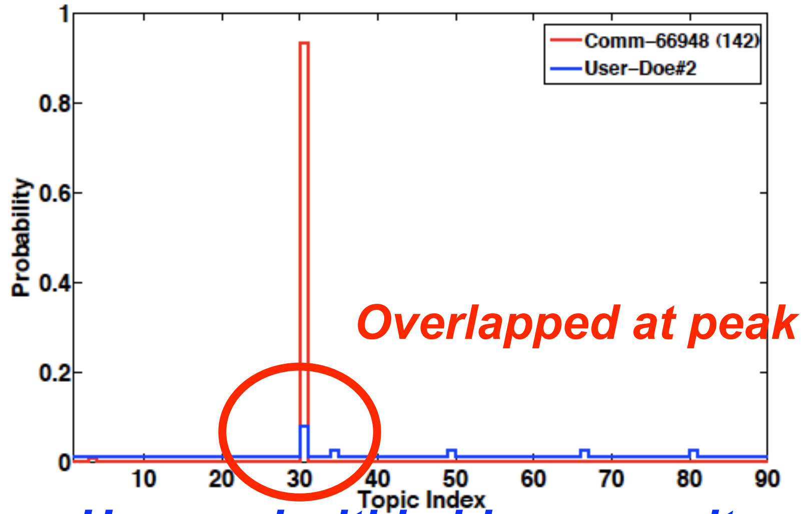
Scattered topic dist.

Analysis of Latent Information from LDA (cont.)

User1 whom LDA ranks better



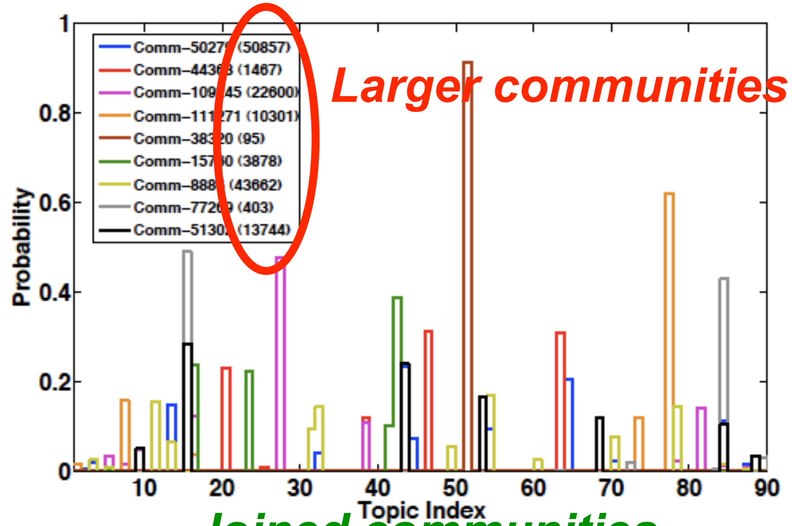
Joined communities



Overlapped at peak

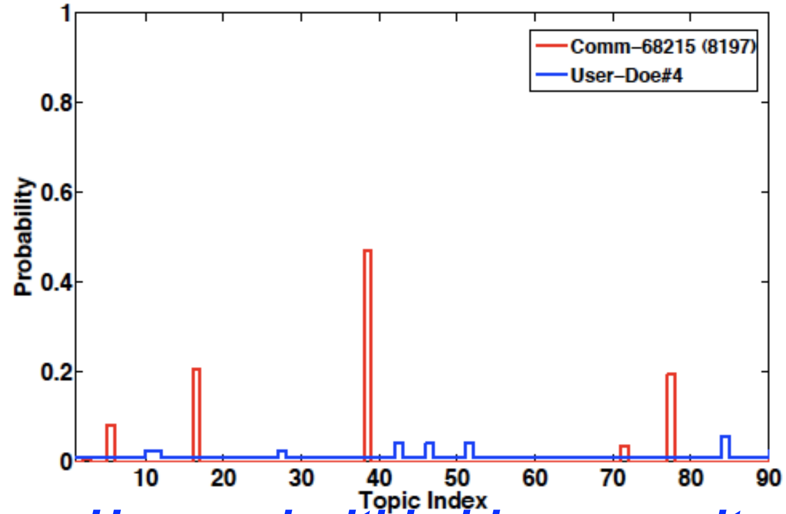
User and withheld community

User2 whom ARM ranks better



Larger communities

Joined communities



User and withheld community

Runtime Speedup of parallel LDA

Runtime for LDA using different number of machines

- Use up to 32 machines
- 150 topics, 500 iterations
- Reduce time from 8 hrs to 45 mins

Machines	Comp	Comm	Sync	Total	Speedup
1	28911s	0s	0s	28911s	1
2	14543s	417s	1s	14961s	1.93
4	7755s	686s	1s	8442s	3.42
8	4560s	949s	2s	5511s	5.25
16	2840s	1040s	1s	3881s	7.45
32	1553s	1158s	2s	2713s	10.66

Linear speedup

- When increasing the # of machines
  - Computation time was halved
  - Communication time increased
  - Communication has larger impact on speedup

Conclusions

Discovery of user latent behavior on Orkut

- Compared ARM and LDA for community recommendation task
  - Used top-*k* ranking metric
- Analyzed latent information learned from LDA
- Parallelized LDA to deal with large data

Future work

- Extend LDA method to consider the strength of relationship between a user and a community
- Extend ARM method to take multi-order rules into consideration

Parallel LDA code release

- http://code.google.com/p/plda/ (MPI implementation)