LIKE and Recommendation in Social Media

Dongwon Lee[#] and Huan Liu[*]

[#]The Pennsylvania State University and National Science Foundation
[*]Data Mining and Machine Learning Lab, Arizona State University

http://goo.gl/Osg0jc

May 18, 2015

Slide Available for Download

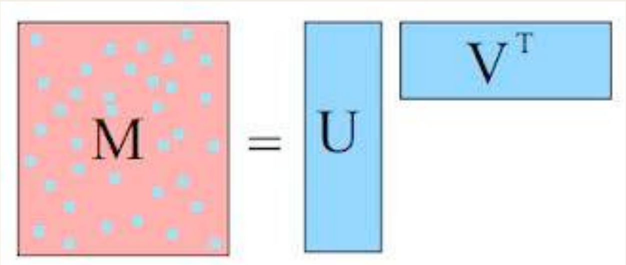
- http://goo.gl/Osg0jc

Like vs. Recommendation

- LIKE: Human-initiated endorsement
  - Eg, I "like" the photo that you posted

Part I

- Recommendation: Machine-initiated endorsement
  - Eg, Amazon "recommends" books that a user may like
  - Machine may use human input: digital footprints



Part II

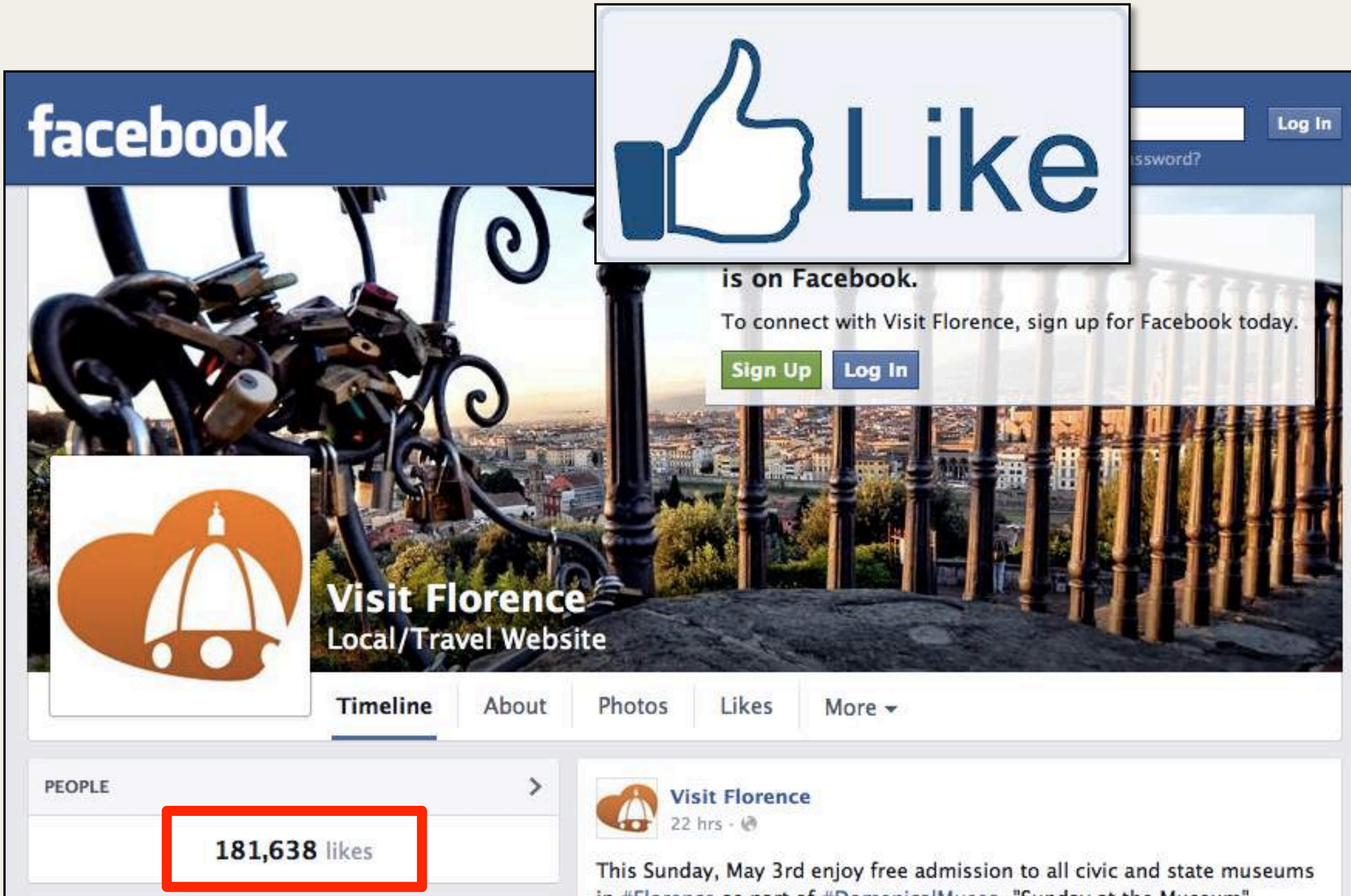
Part 1: LIKE in Social Media

Dongwon Lee

Outline

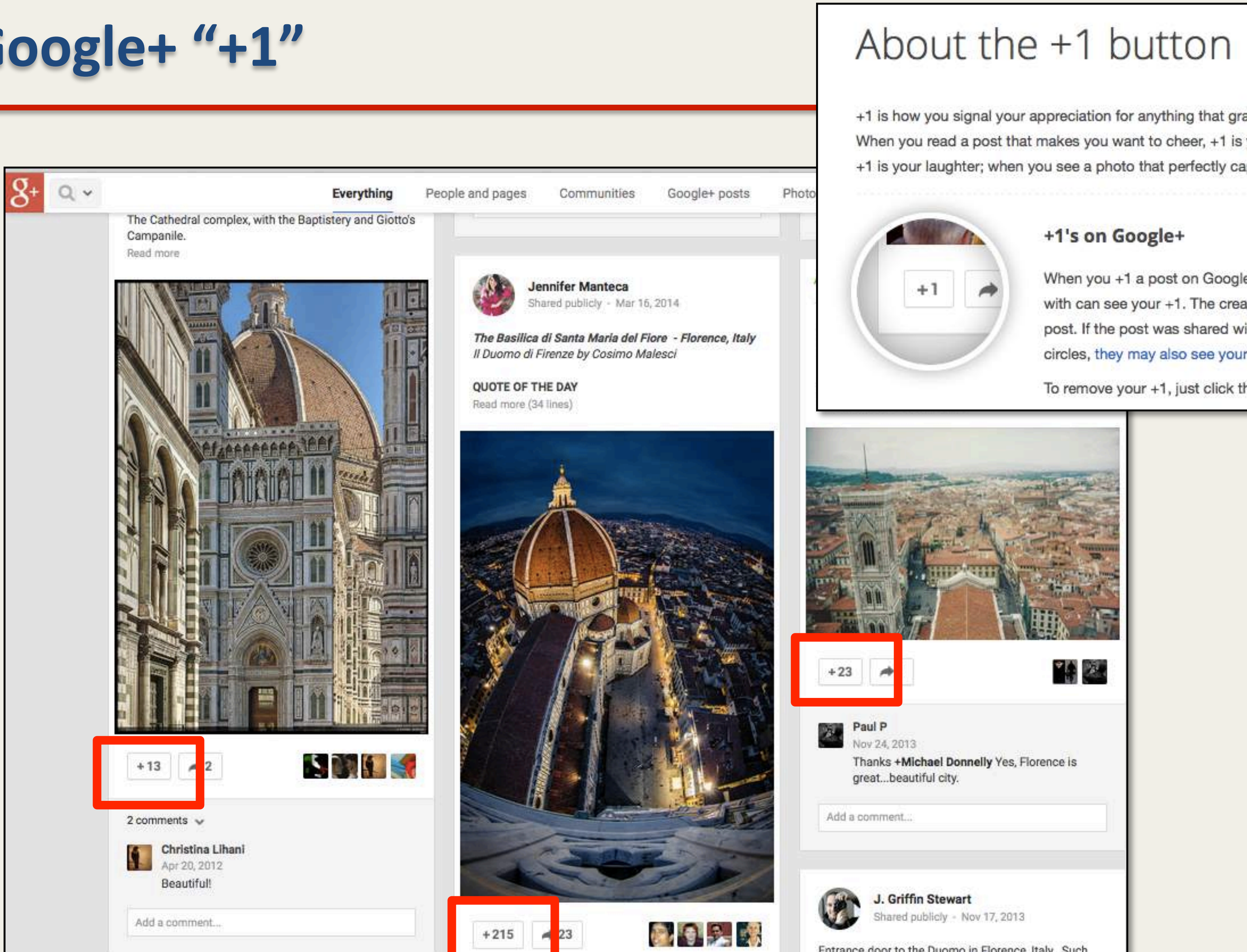
- Introduction
- Understanding LIKEs
- Predicting LIKEs
- Aggregating LIKEs
- Summary

Facebook "Like"



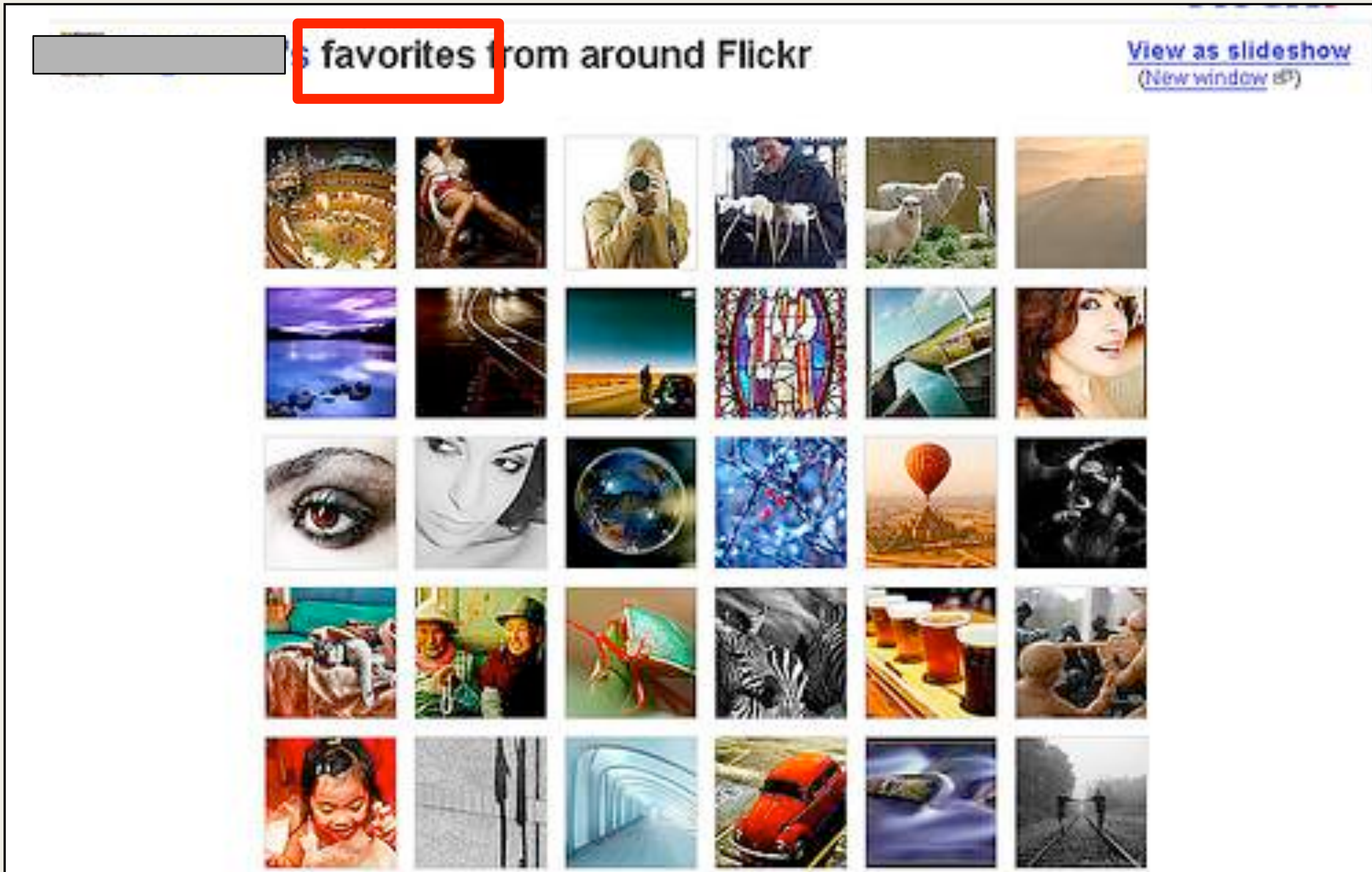
YouTube "Like"

Google+ "+1"

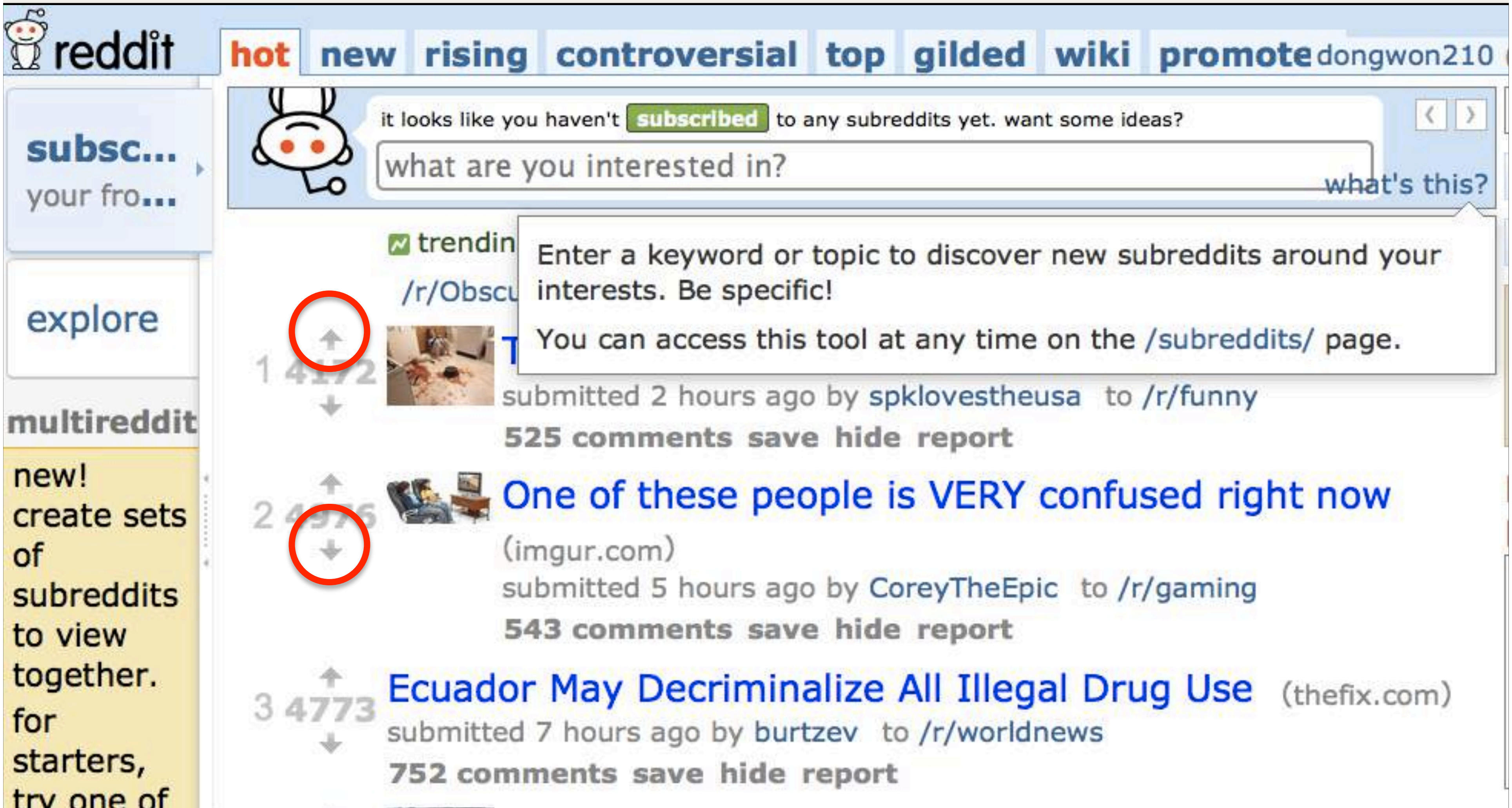

About the +1 button
+1 is how you signal your appreciation for anything that gra
When you read a post that makes you want to cheer, +1 is y
+1 is your laughter; when you see a photo that perfectly cap
+1's on Google+
When you +1 a post on Google
with can see your +1. The creat
post. If the post was shared wit
circles, they may also see your
To remove your +1, just click th
Everything
People and pages
Communities
Google+ posts
Photo
The Cathedral complex, with the Baptistery and Giotto's Campanile.
Read more
+13
2
2 comments
Christina Lihani
Apr 20, 2012
Beautiful!
Add a comment...
Jennifer Manteca
Shared publicly · Mar 16, 2014
The Basilica di Santa Maria del Fiore - Florence, Italy
Il Duomo di Firenze by Cosimo Malesci
QUOTE OF THE DAY
Read more (34 lines)
+215
23
+23
Paul P
Nov 24, 2013
Thanks +Michael Donnelly Yes, Florence is great...beautiful city.
Add a comment...
J. Griffin Stewart
Shared publicly · Nov 17, 2013
Entrance door to the Duomo in Florence, Italy. Such

Twitter "Favorites"

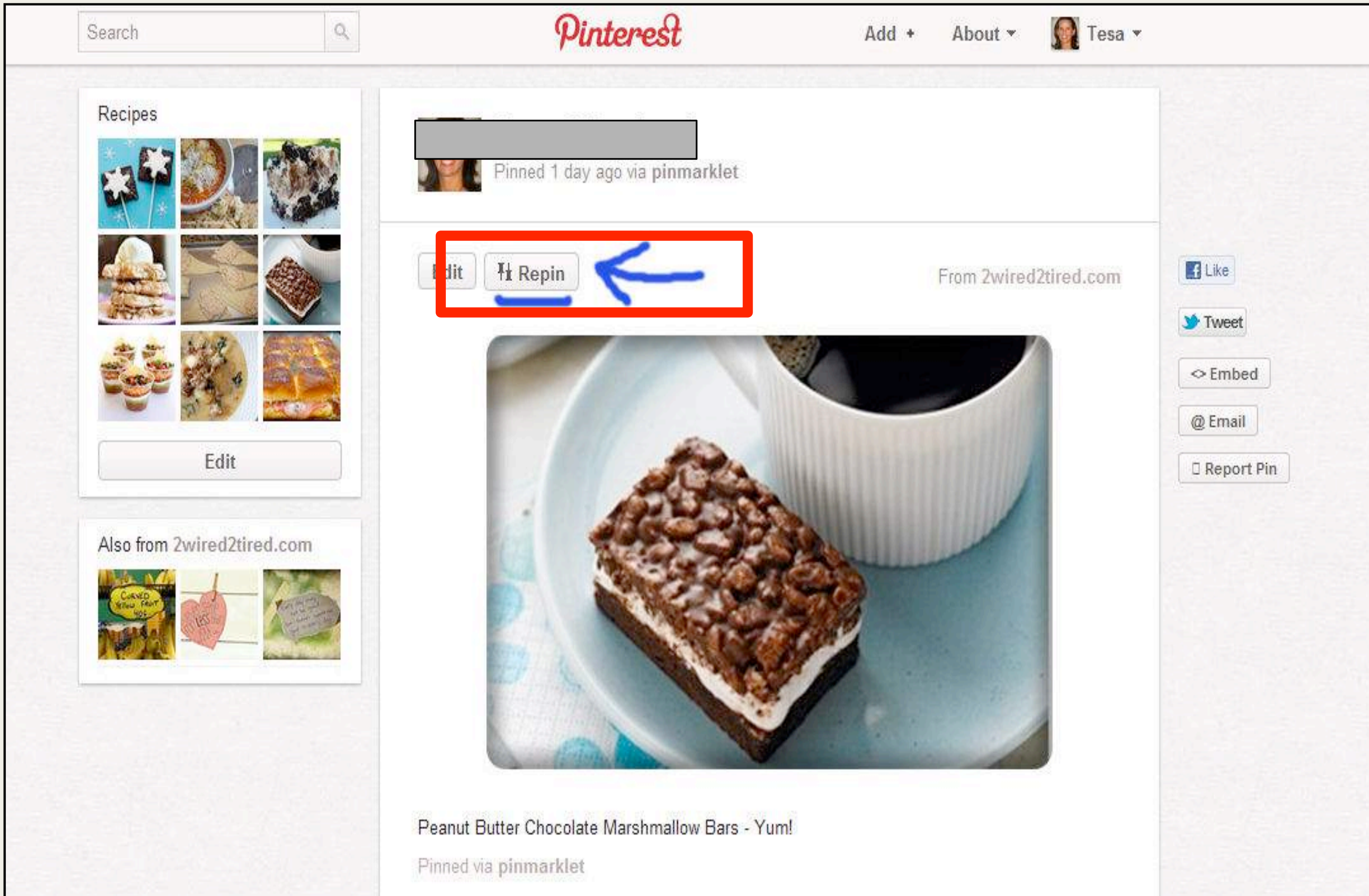
Flickr “Favorites”



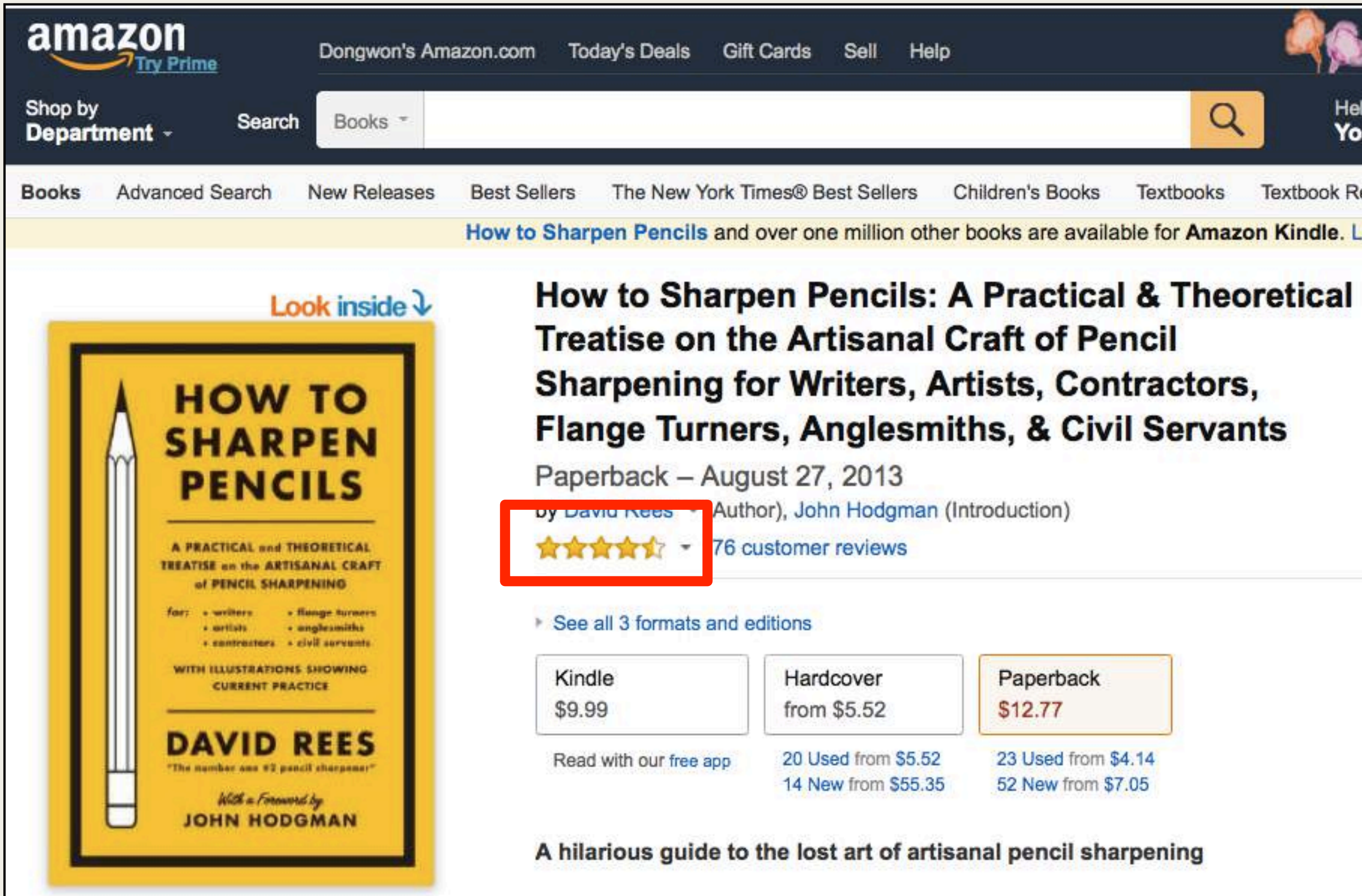
Reddit "Upvote"



Pinterest "Re-pin"

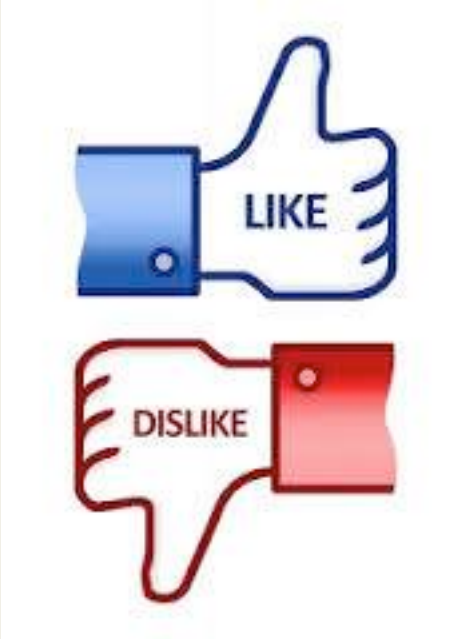


Amazon Star Ratings



Values of LIKEs

- Binary: {0, 1}
  – Facebook LIKE
  – Google+ +1
  – Flickr favorite
- Ternary
  – Reddit Votes: {+1, -1, 0}
  – YouTube {Like, Dislike, None}
- *N*-ary
  – Amazon book ratings [1..10]
  – YouTube "Like" used to be [1..5] (until 2010) then changed to ternary



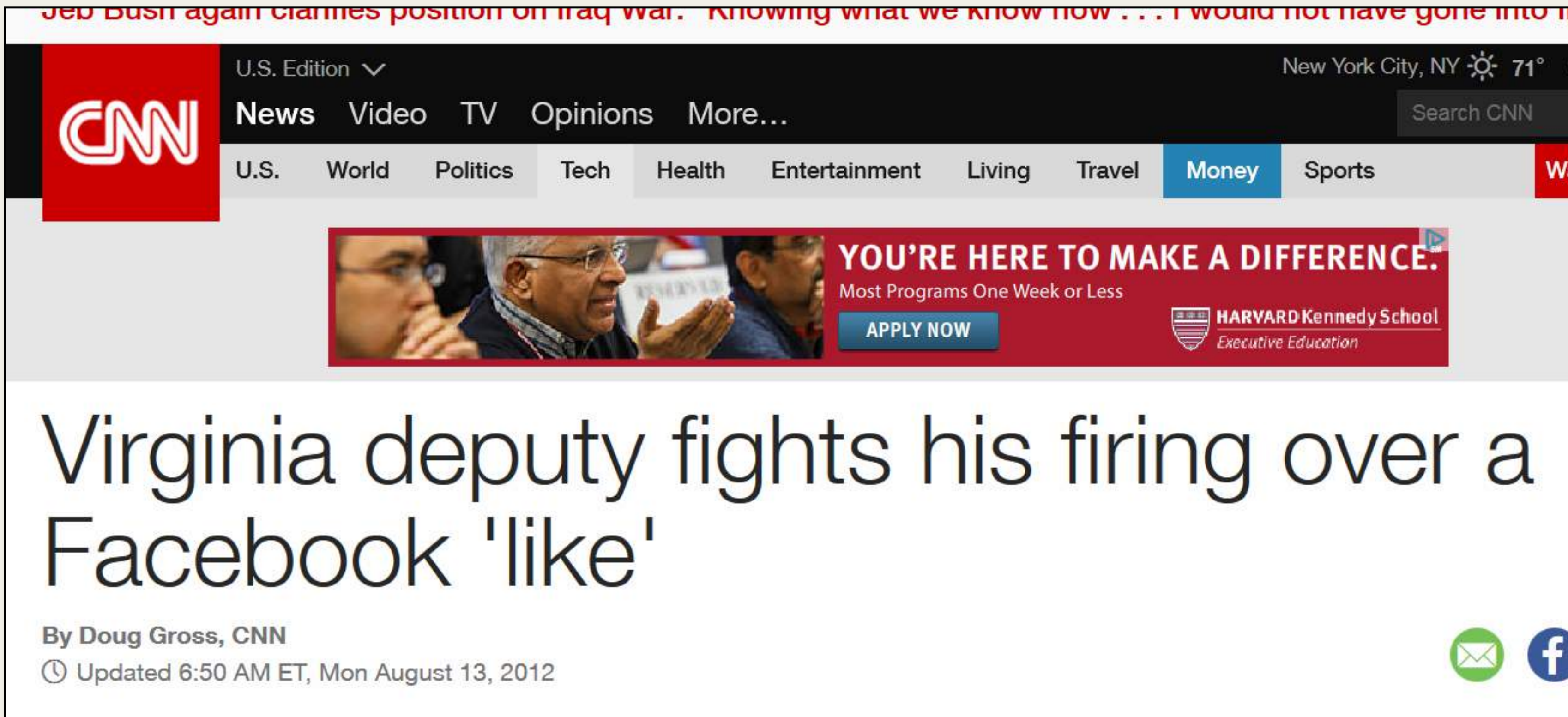
Meanings of LIKEs

- I saw it or I was here
- Preference
- Endorsement of taste/vote
- Fan or advocate
- Agreement
- Subscription
- Self-expression
- Reciprocal relationship
  – Eg, you liked my photo, so I like your photo

- An interesting RQ itself !

Meaning of LIKEs

- Subscription vs. Endorsement
- LIKE should be protected by the First Amendment right to free speech? **[Robbins, 2013]**



Objectives

- Understand users in social media through the lens of LIKE
  - LIKE based network analysis
  - LIKE as a dependent factor for analysis
  - LIKE as a feature for machine learning
- Understand the evolution of LIKE
  - Predict # of LIKE and LIKE relationship
- Use LIKE for recommendation
  - Aggregate LIKE

Outline

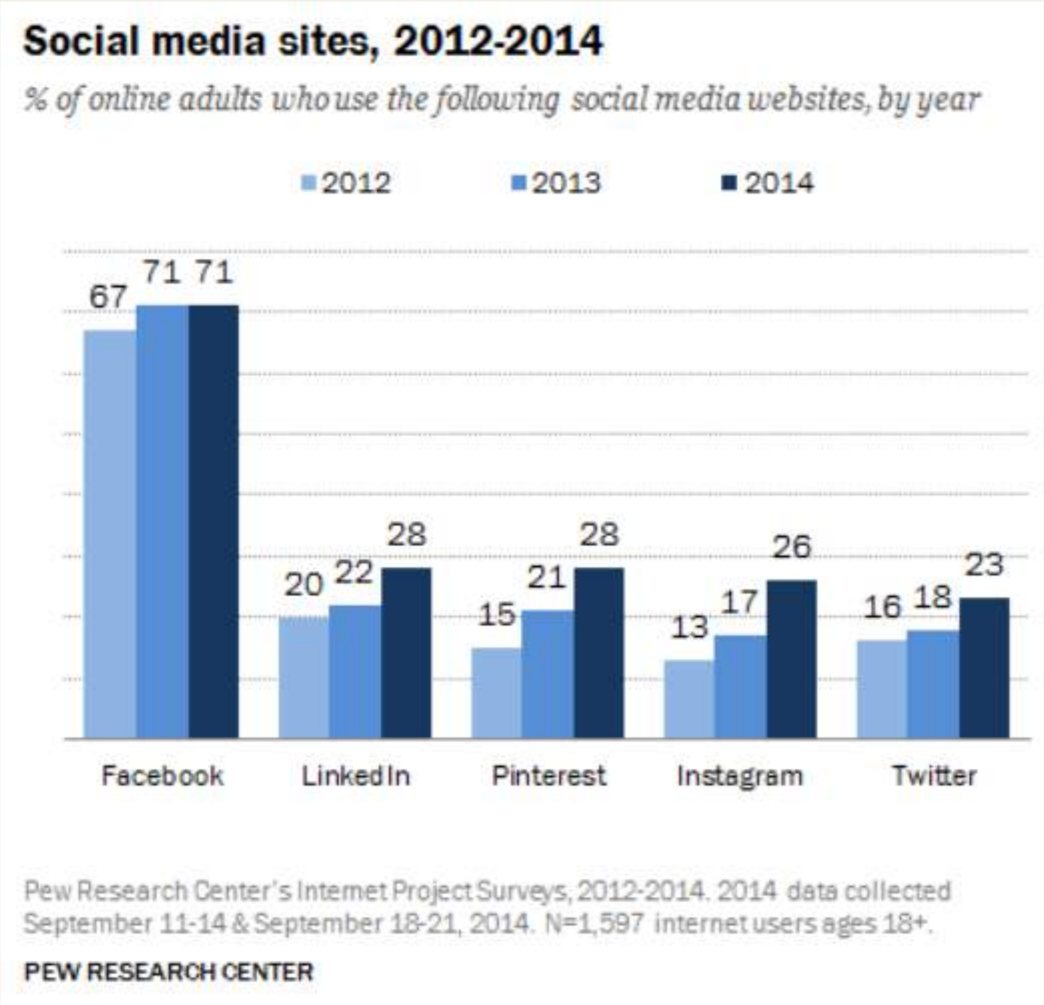
- Introduction
- Understanding LIKEs
- Predicting LIKEs
- Aggregating LIKEs
- Summary

LIKE Activities

- Indication of one's shared interests toward the content or the original poster
- May lead to creating & fostering social relationship
- Commercial values
  - Providing recommendations of products, users, or activities
- RQ: Study SNs through the lens of LIKEs
  - Structure
  - Context
  - Influence

Medium to Study

- Pew research survey, 2014



Medium to Study

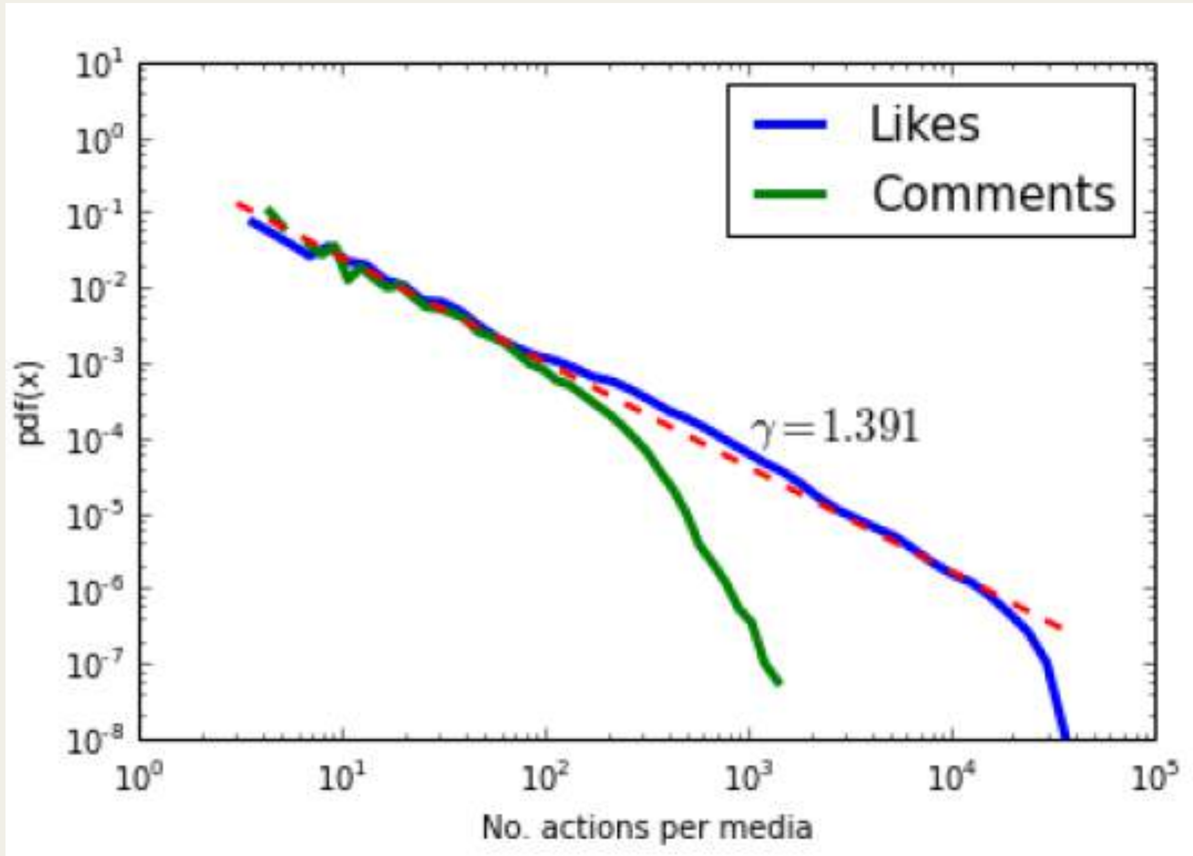
- *Instagram*: an online photo-sharing service
  - Enables users to take picture and videos, apply digital filter to them, and share them on a variety of SNSs
- Popularity
  - 2/2013: 100 million active users
  - 9/2013: 150+ million monthly active users
  - Very popular on mobile platform

LIKE Distribution [Ferrara et al., 2014]

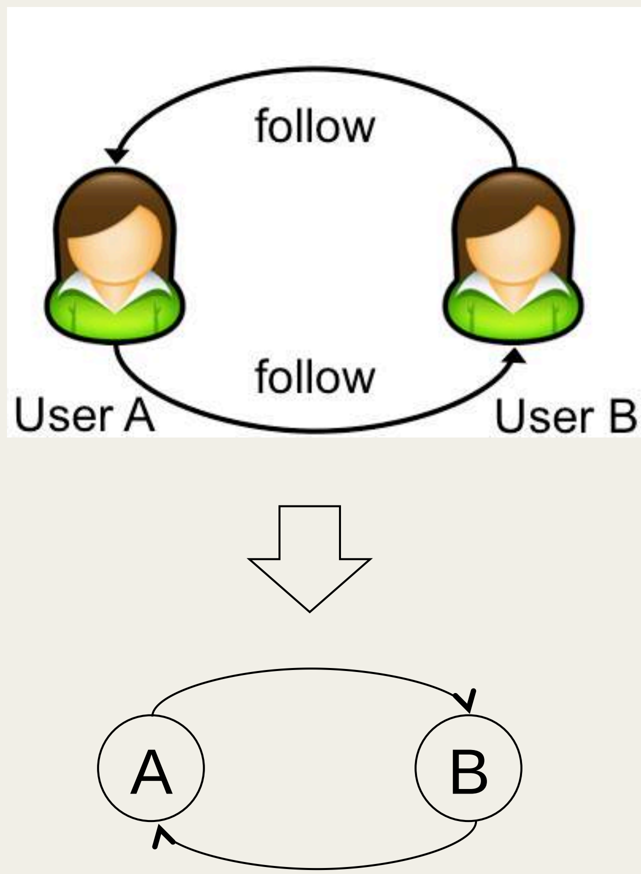
- LIKE and comments show different behaviors
- Different cost between LIKE and comments

1-month
Instagram data

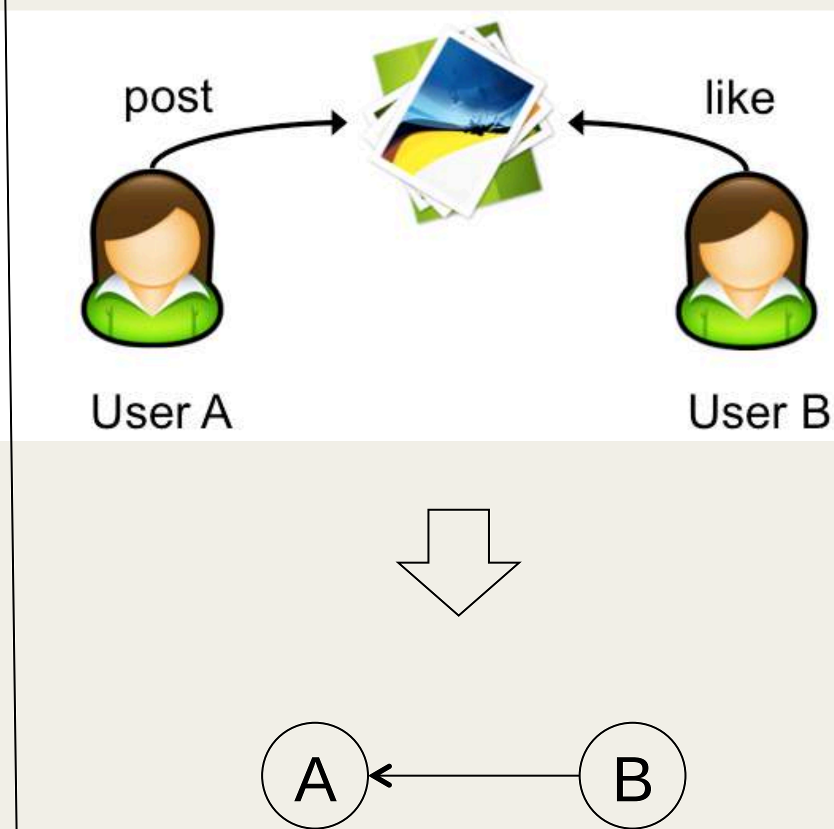
Follow Network (FN)
2K users
1.7M photos
1.2B LIKEs
41M comments



LIKE Network [Jang et al., TR]



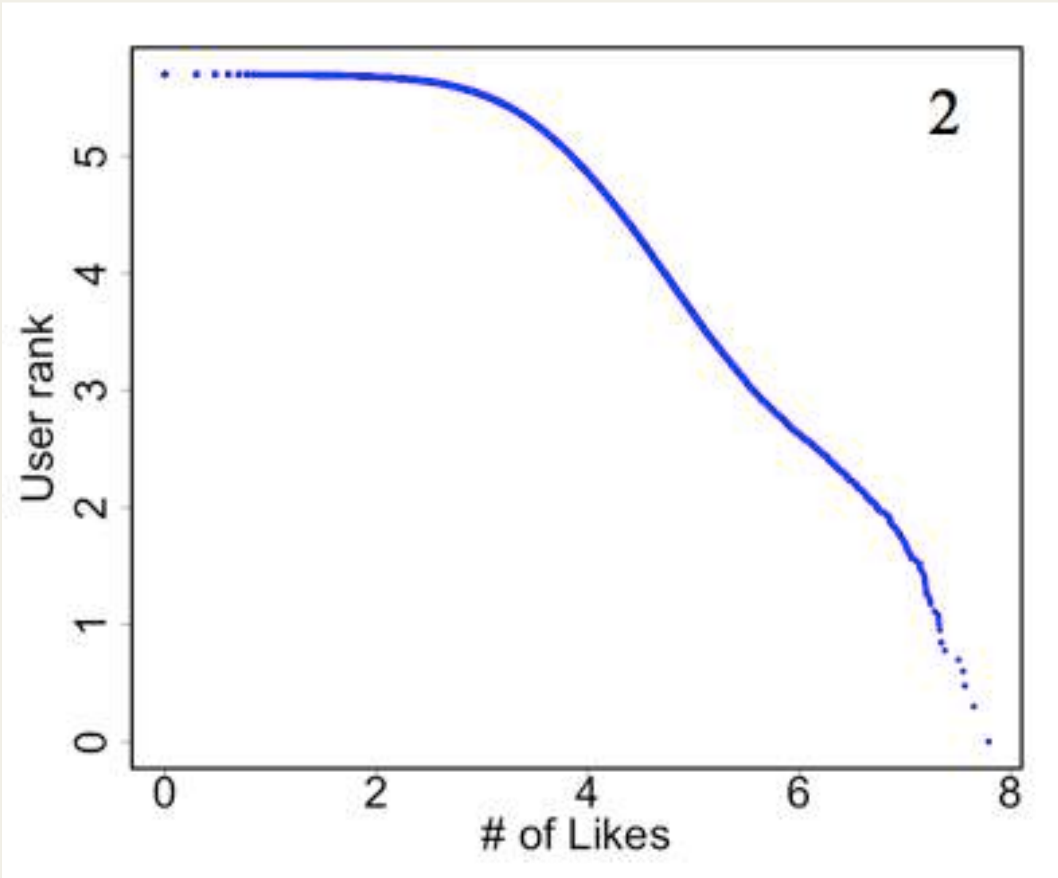
Friend Network (FN)



LIKE Network (LN)

LIKE Network [Jang et al., TR]

- 1K seed users, 20M users, 2B LIKEs
- On average, 55.6% LIKEs are from followers



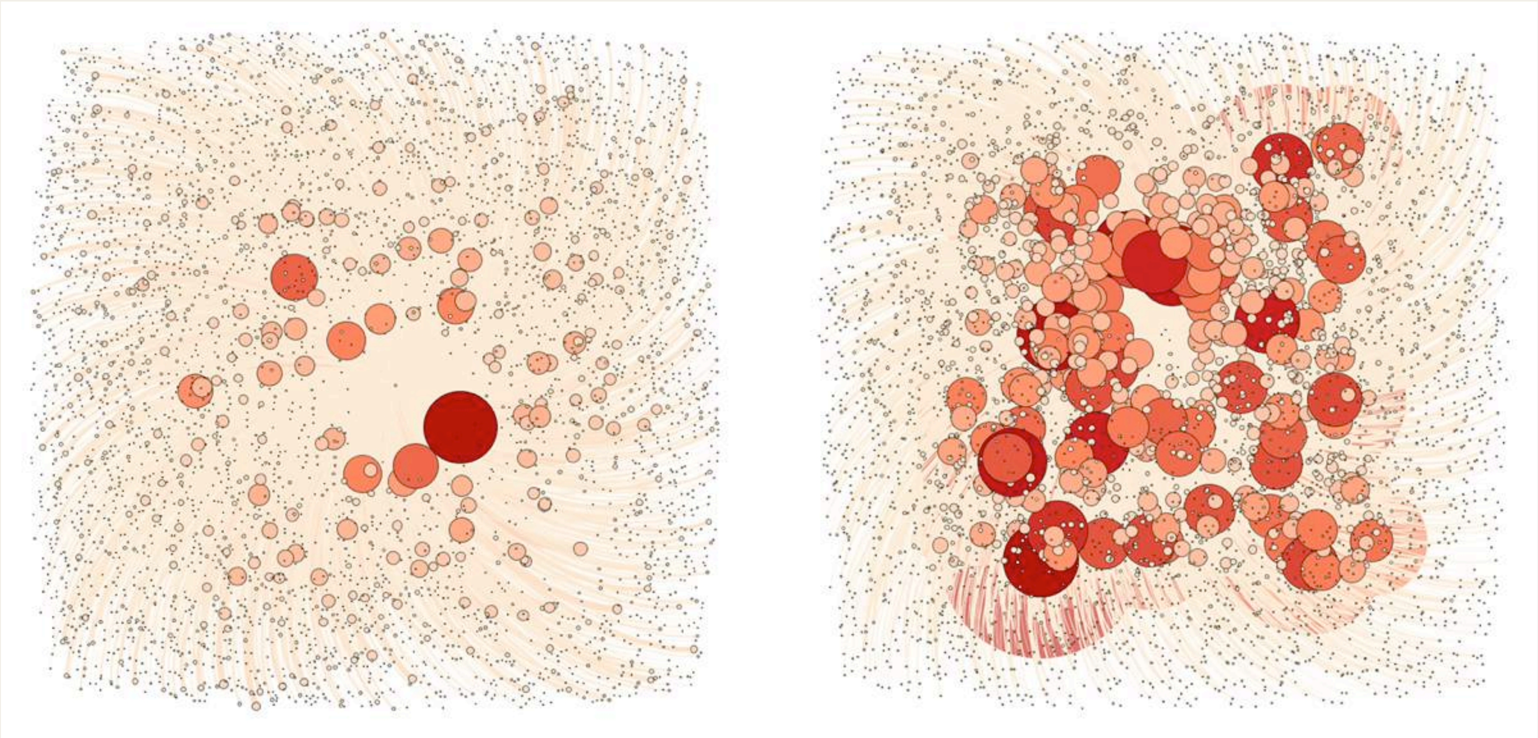
LIKE Network [Jang et al., TR]

- Descriptive statistics per user (N=500K)

Variable	Median	Mean	Max	S.D.
# Photos	166	309	57,925	485
# Likes	1,984	11,240	61,606,804	235,974
# Tags	103	227	97,248	1,028
# Comments	58	321	1,112,862	3,851
# Followers	624	2,403	2,751,722	16,454
# Follows	292	723	5,291,779	18,305

LIKE Network [Jang et al., TR]

- 2 Examples of LNs of random posters p_1 and p_2
  - Most LIKEs were from users who gave a single Like
  - High sparseness of the network



10K Likes, 34% from followers

63K Likes, 56% from followers

Contexts and LIKE Activities [Jang et al., TR]

- Instagram does not have genres
- LDA-based topic generation
- Top-100 topics from Mallet
- Bottom-up semi-manual construction

- 3rd parties
  - Eg, tagsforlikes.com, tagstagram.com

Contexts and LIKE Activities [Jang et al., TR]

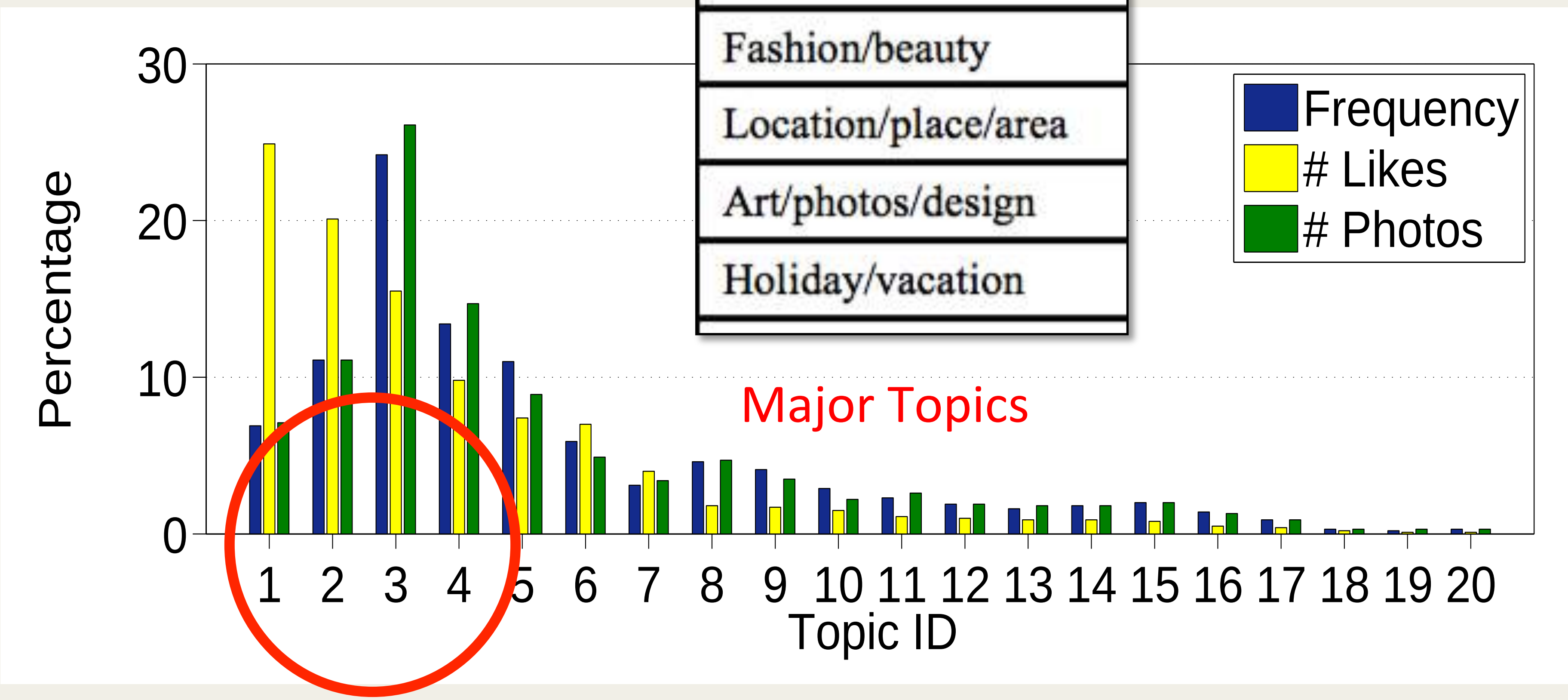
- Top-20 topics

ID	Topic	Tag examples
1	Nature	sky, nature, flowers, sea
2	Fashion/beauty	makeup, jewelry, model
3	Location/place/area	nyc, boston, spain, Italy
4	Art/photos/design	photo, interior, architect
5	Holiday/vacation	party, holiday, vacation
6	Mood/emotion	love, cute, happy, smile, great
7	Social/people/family	family, girlfriend, gay, folks
8	Sports/activity	skateboarding, hiking, soccer
9	Entertainment	music, movie, pop, rock
10	Follow/shoutout/like	tagsforlike, followme, likes

Contexts and LIKE Activities [Jang et al., TR]

- Ratio of frequency, # of *Likes*, and # of photos for the topics

Topic
Nature
Fashion/beauty
Location/place/area
Art/photos/design
Holiday/vacation

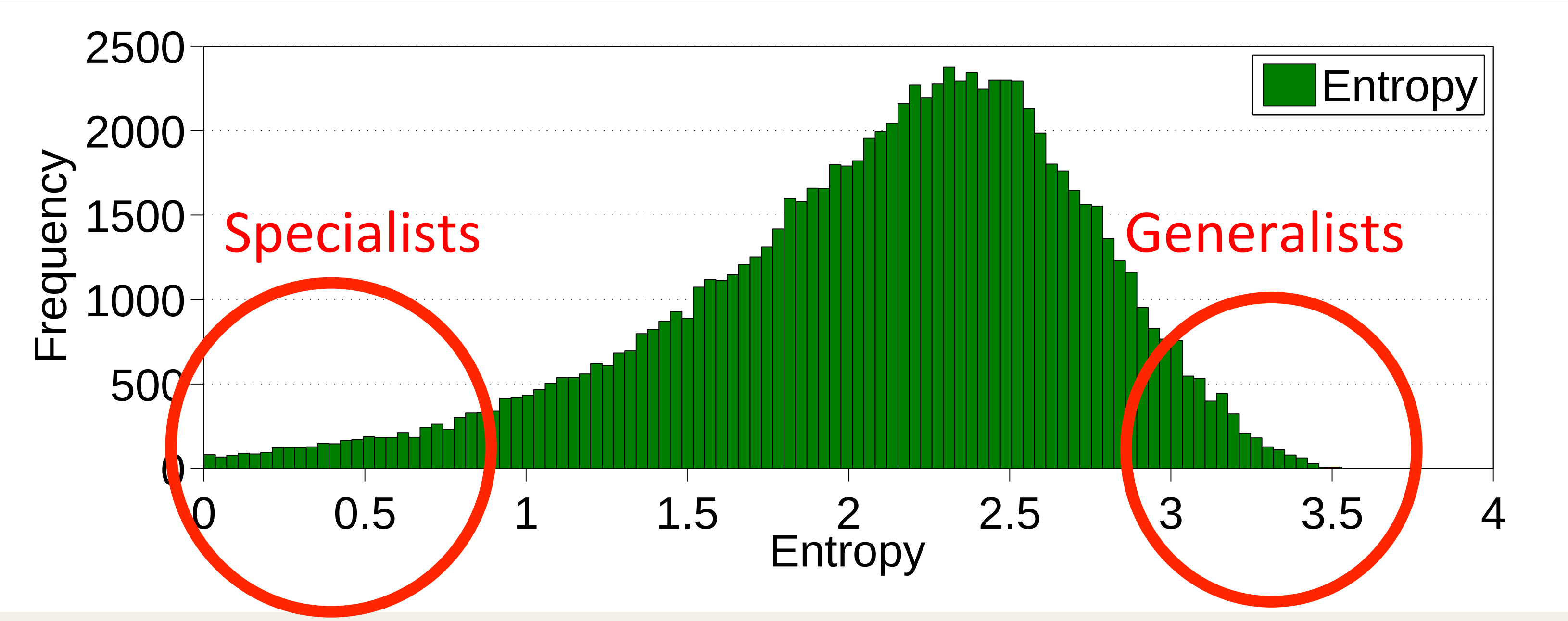


Contexts and LIKE Activities [Jang et al., TR]

- A measure of the uncertainty in a random variable

$$Entropy(p) = -\sum_{i=1}^{20} P(x_i) \log P(x_i)$$

- High/low entropy ➔ diverse/specific topic

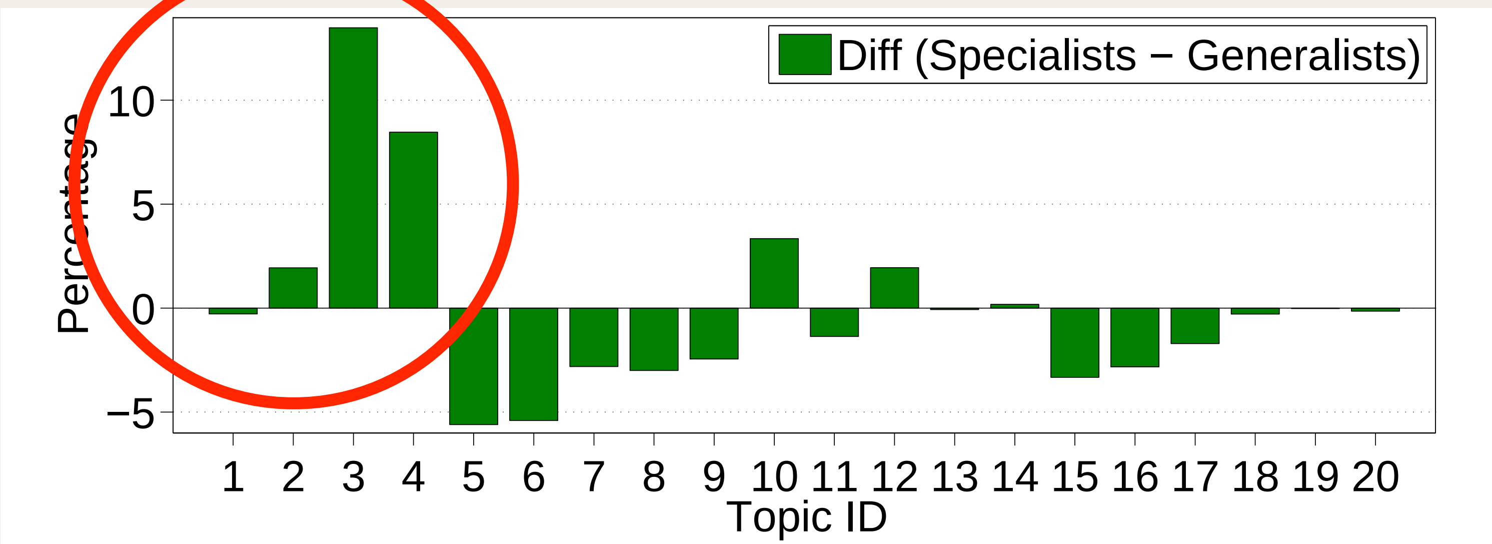


Contexts and LIKE Activities [Jang et al., TR]

- Specialists vs. Generalists

Type	# Posters	# Likes		
		Mean	Median	S.D.
Specialists	5,594	101,666	10,893	1,358,885
Generalists	3,230	15,989	2,375	367,911

- Topic distribution



Popularity and LIKE [Jang et al., TR]

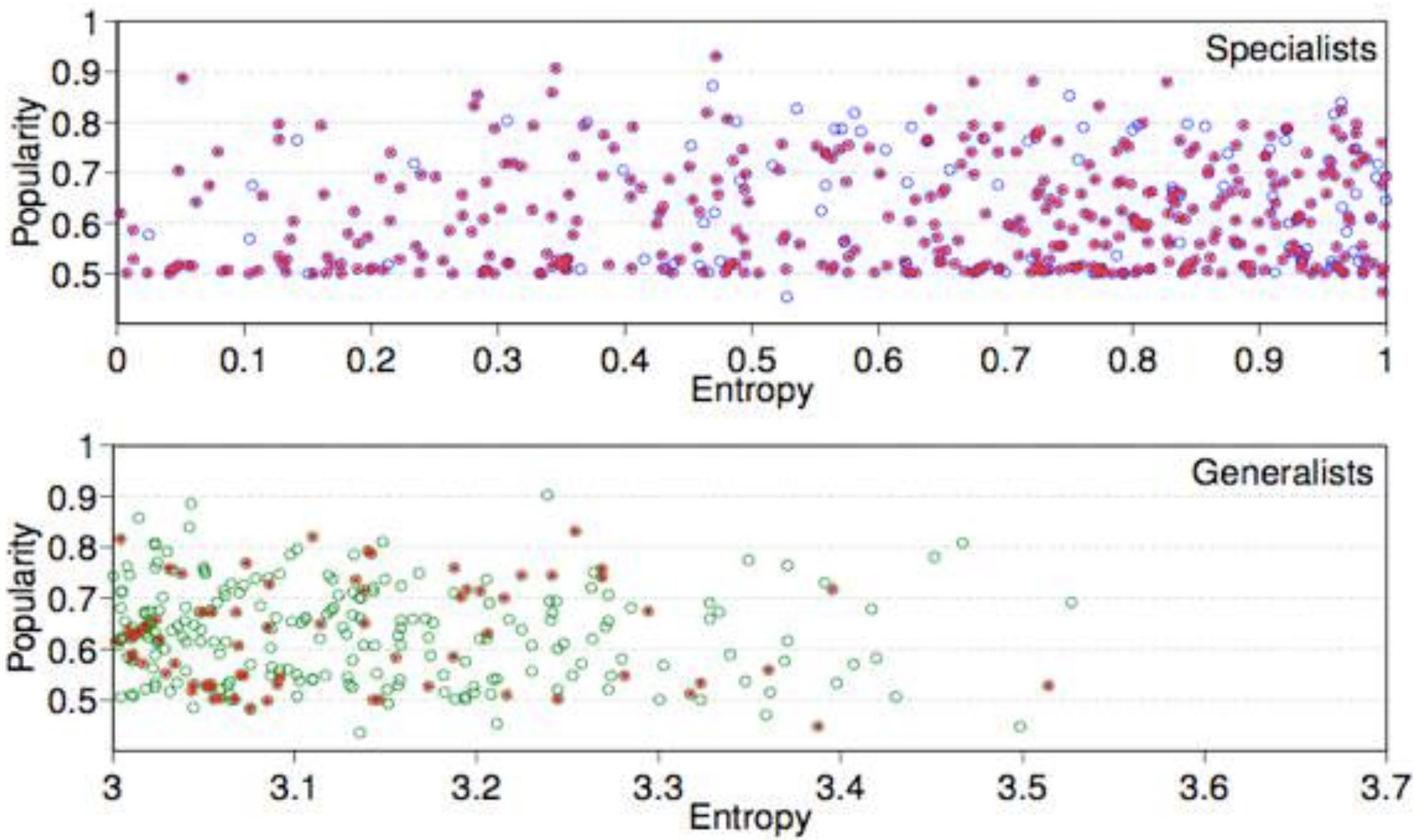
- A poster p_1 is more popular than a poster p_2 if:
  - p_1 receives on average a more number of LIKEs per photo than p_2 does, and
  - p_1 has a higher ratio of followers over follows than p_2 has

$$Popularity\ (p) = \alpha \left(\frac{\frac{L}{P}}{\sqrt{1+\left(\frac{L}{P}\right)^2}} \right) + (1-\alpha)\left(\frac{F}{F+F'}\right)$$

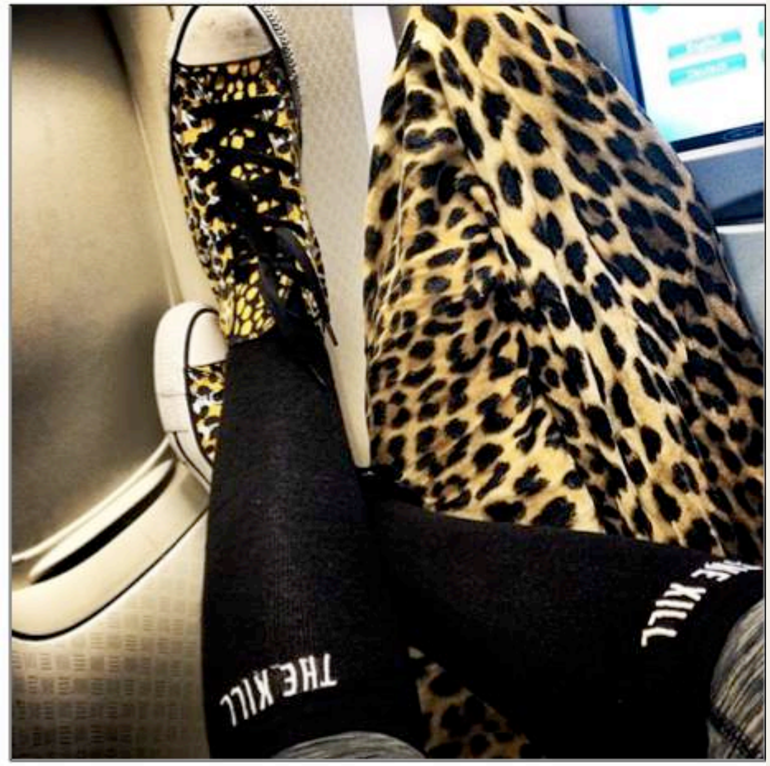
- L: # of LIKEs
- P: # of photos

F: # of followers
F': # of follows

Popularity and LIKE [Jang et al., TR]



Popularity and LIKE [Jang et al., TR]

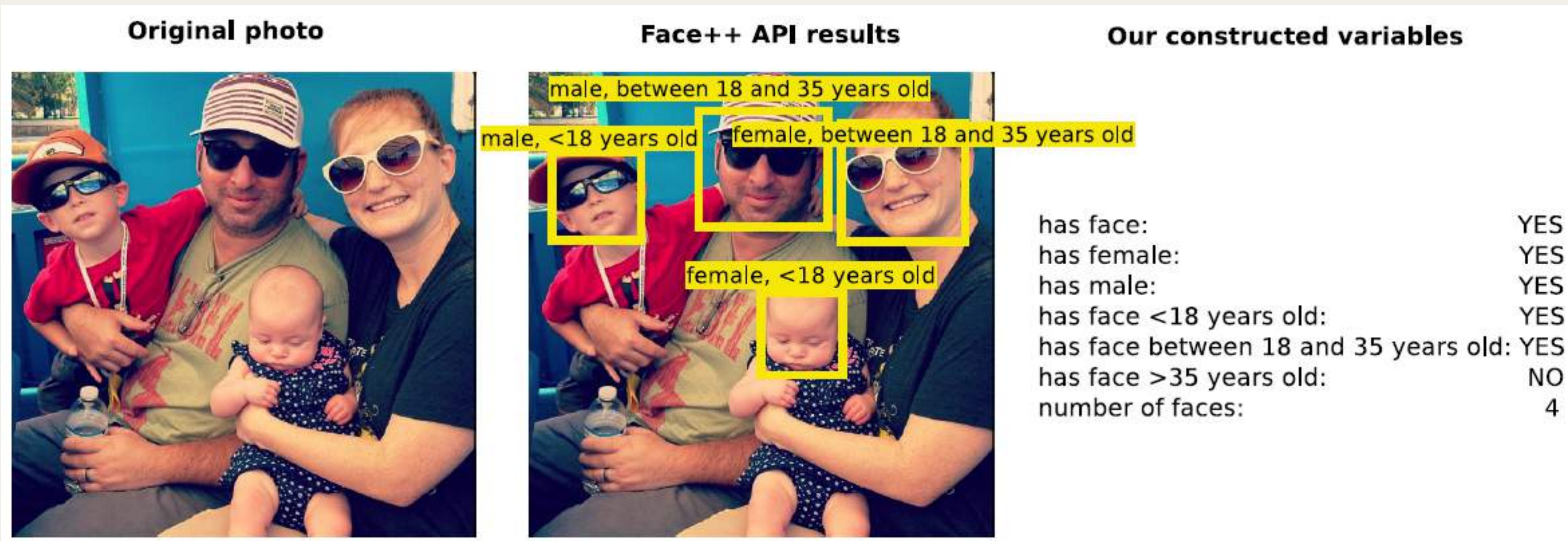


Popular Specialist
(P > 0.7, LIKE = 6K)

Popular Generalist
(P > 0.7, LIKE = 15.2K)

Faces Engage More LIKEs [Bakhshi et al., 2014]

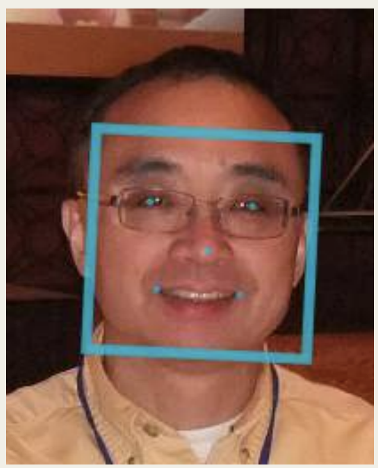
- 1M Instagram dataset
- 38% more LIKEs with faces
- 32% more comments with faces

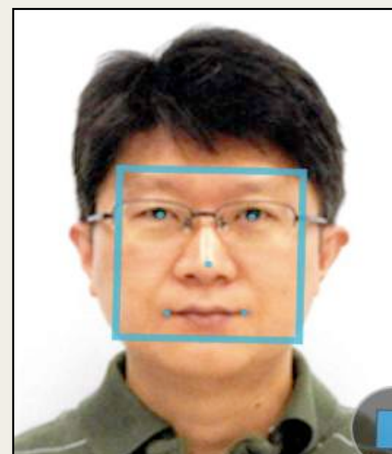


Faces Engage More LIKEs [Bakhshi et al., 2014]

Validation test	Accuracy	Margin of error
has face	97%	0.75%
has female face	96%	0.86%
has male face	96%	0.86%
has face $<$ 18 years old	93%	1.11%
has face between 18 and 35 years old	96%	0.86%
has face $>$ 35 years old	99%	0.44%

http://www.faceplusplus.com/demo-detect/



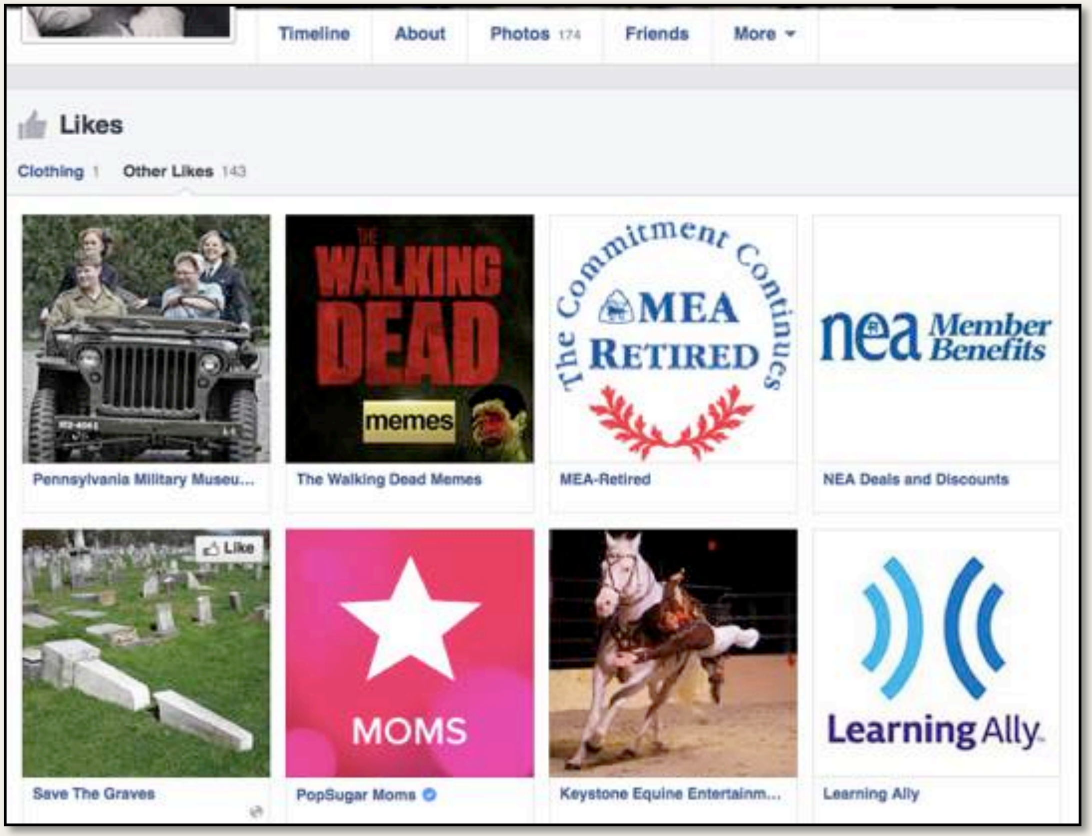
10 - 47

10 – 28

You Are What You LIKE

- *Hypothesis*: The LIKE pattern in social media may be correlated with users' personal traits

Facebook LIKEs



You Are What You LIKE

Part 2: Recommender Systems

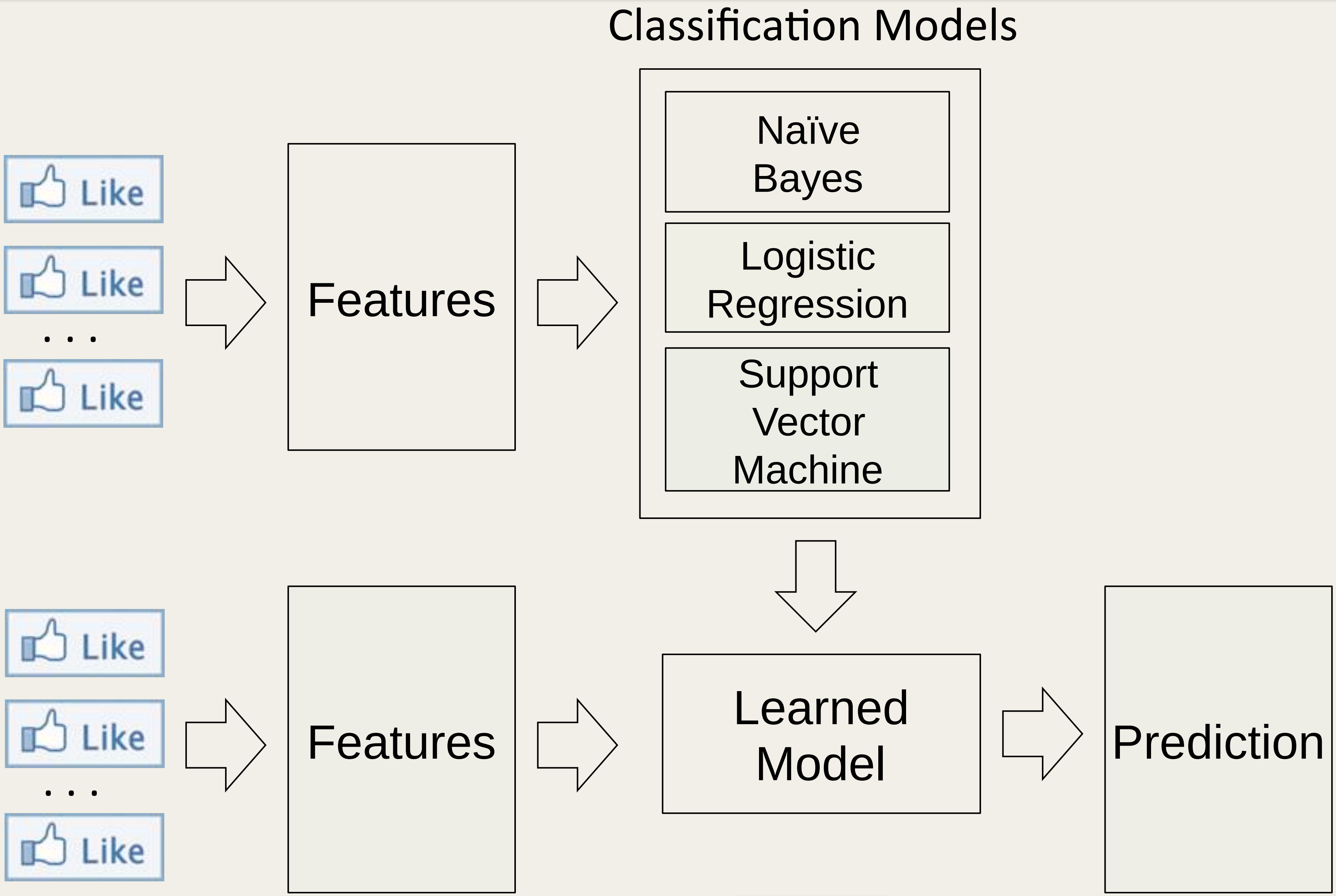
Items to LIKE

Users

	1	2	3	4	5	6	…	N
1	Like		Like			?		
2				Like				Like
3			Like	Like				
?				Like				Like
…								
M	Like					Like		Like

The problem of Inferring about Users

Workflow



Classification Models
Like
Like
. . .
Like
Features
Naïve Bayes
Logistic Regression
Support Vector Machine
Like
Like
. . .
Like
Features
Learned Model
Prediction

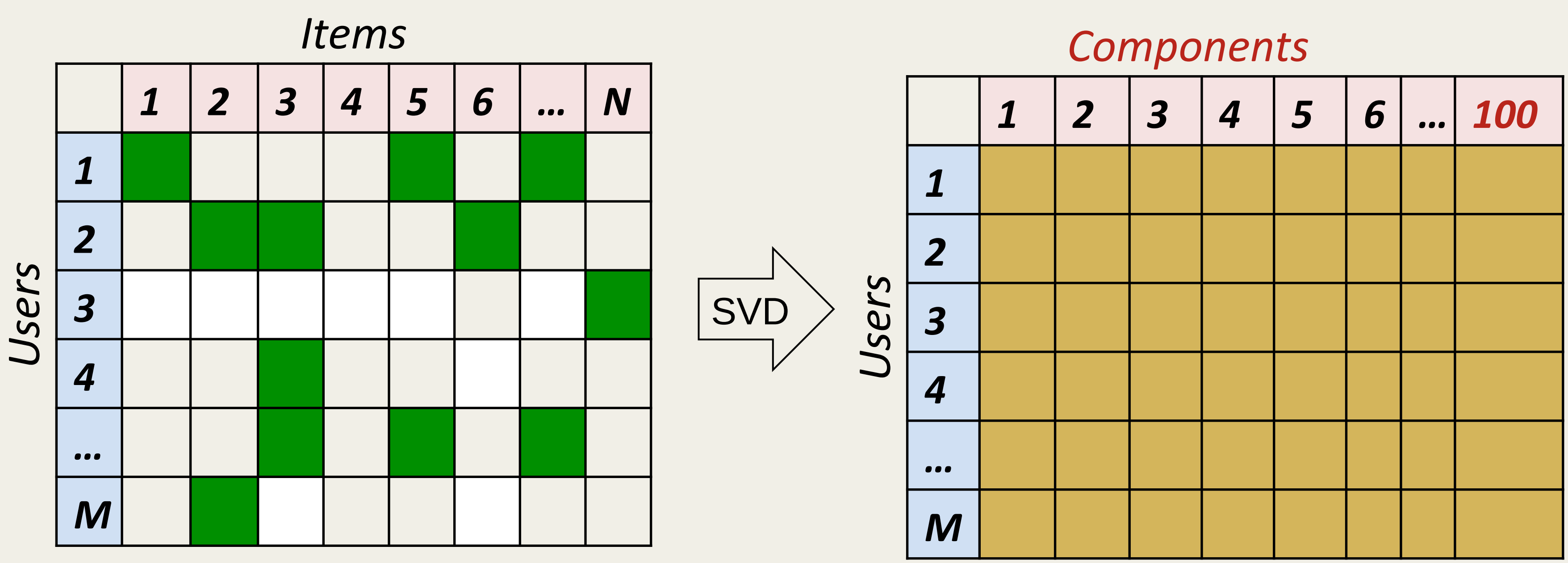
Features from LIKE: Categories

- Semantic labels
- Reduced dimension

Example Like item	Category
Nintendo NES	Games/toys
New Orleans Ice Cream	Food/beverages
Barack Obama	Politician
The Bold and the Beautiful	TV show
eHarmony	Website
Modern Salon	Health/beauty
Sheraton Abu Dhabi Hotel & Resort	Hotel
Justin Bieber	Musician/band
Mercedes-Benz SLK	Cars
Cosmo's Restaurant & Bar	Restaurant/cafe
Indianapolis Museum of Art	Museum/art gallery
Wheaton Franciscan Healthcare	Health/medical/pharmaceuticals
Chicago Air and Water Show	Attractions/things to do
German Shepherd Pup's	Local business
The British Armed Forces	Aerospace/defense

Features from LIKE: Ratings

- User-Item matrix
- Singular Value Decomposition (SVD): eg, selects top-100 components as features
- Approach by [Kosinski-13]

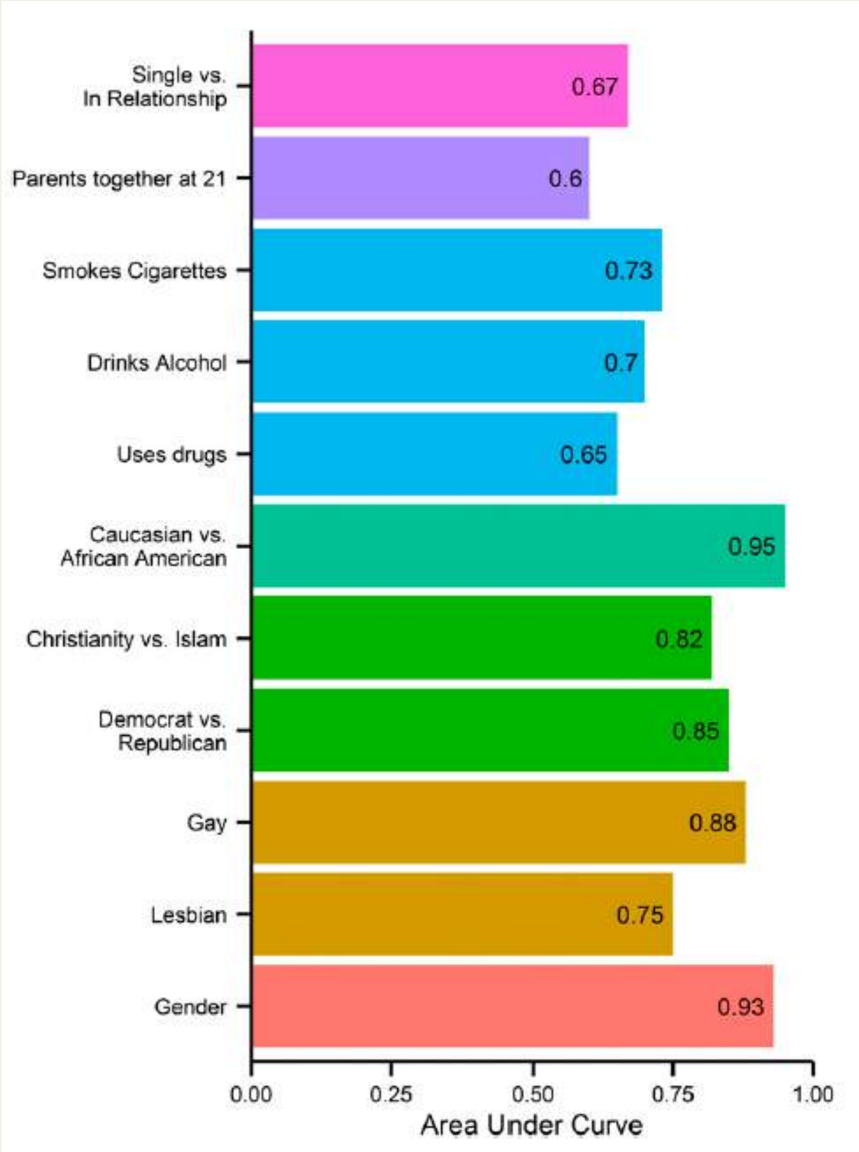


Experimental Set-Up

- Facebook LIKE dataset [Kosinski et al., 2013]
- 1,600 FB users who provided their personal traits information voluntarily

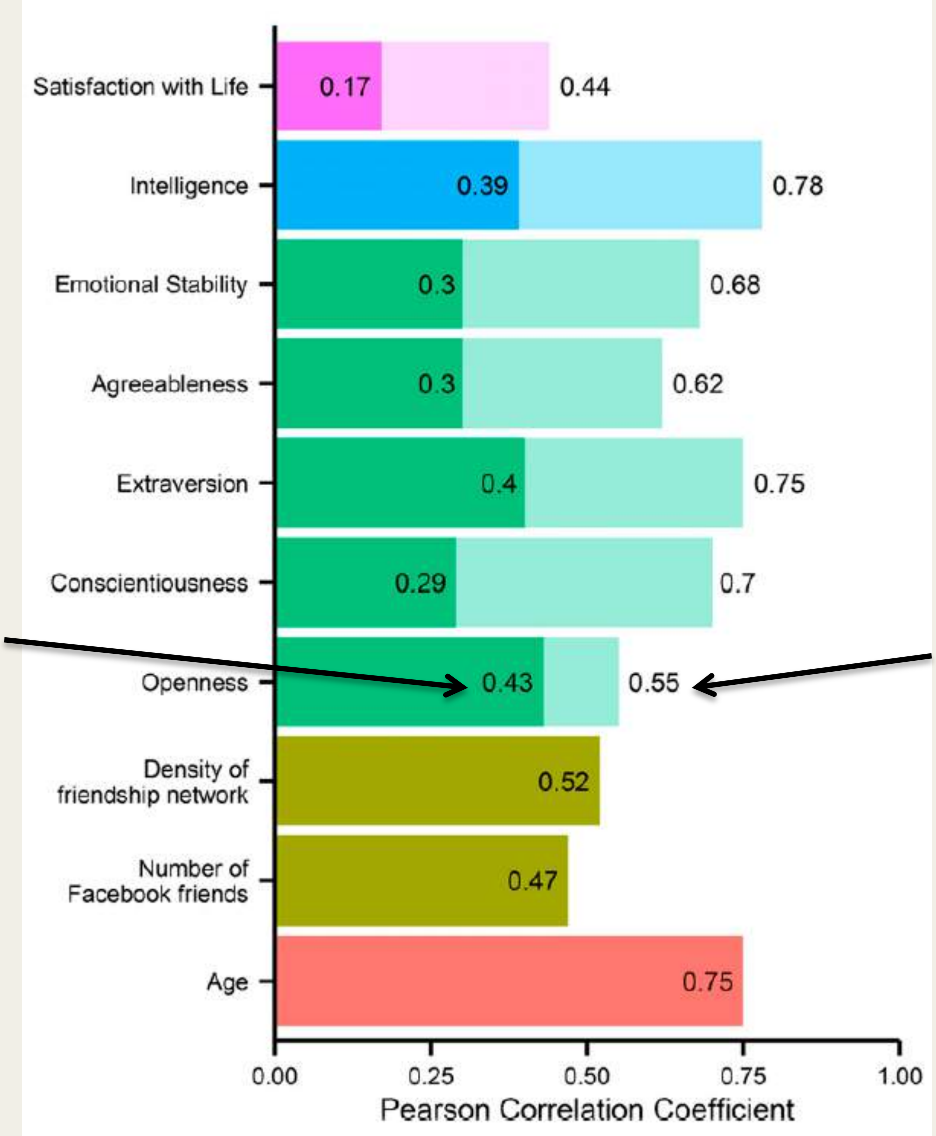
- Gender: {Male, Female}
  – Binary classification
- Age: {20-, 20-30, 30-40, 40-50, 50+}
  – Multi-class classification

Binary Classification Accuracy [Kosinski et al., 2013]



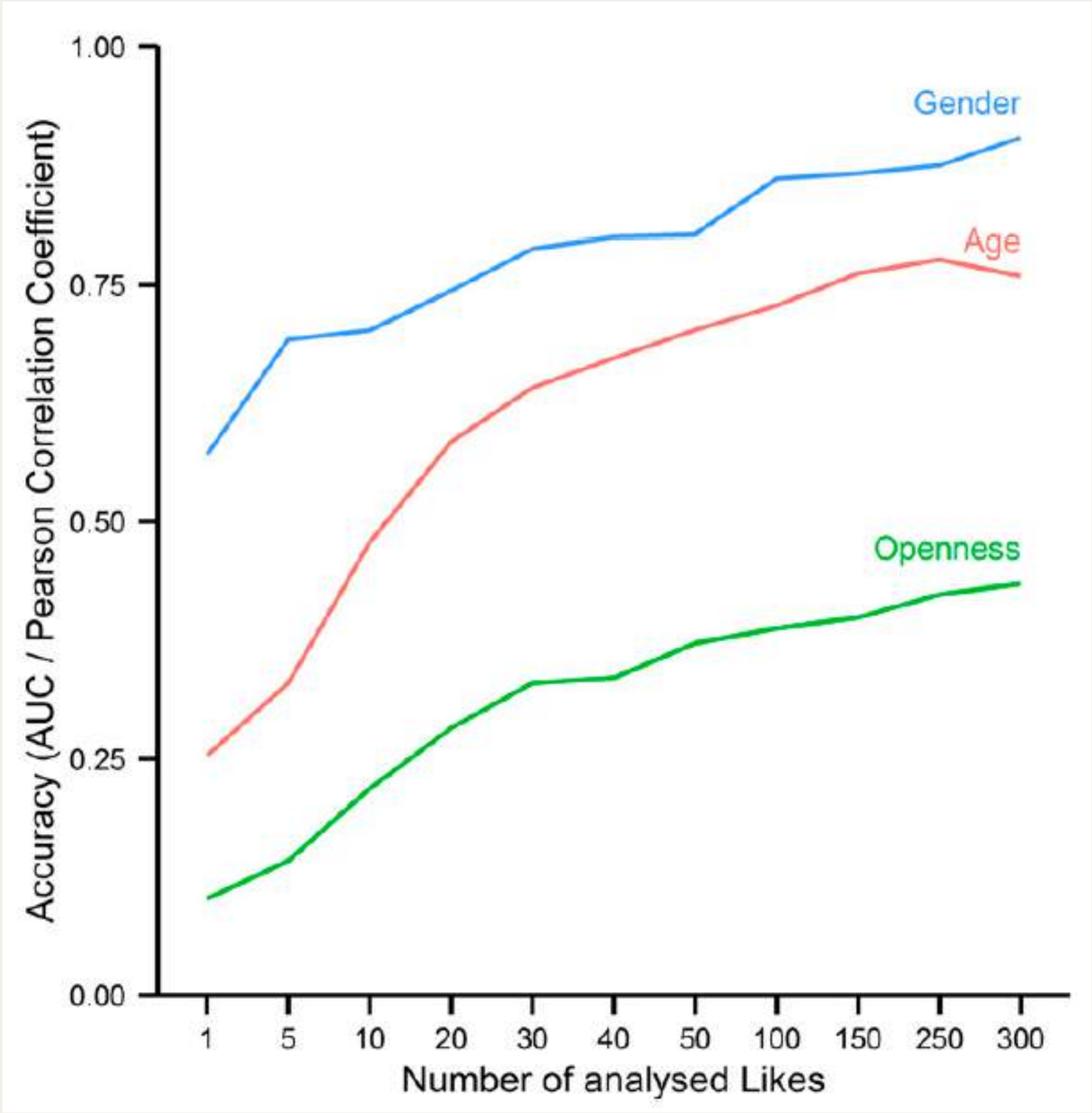
Regression Accuracy [Kosinski et al., 2013]

LIKE-based
SVD prediction



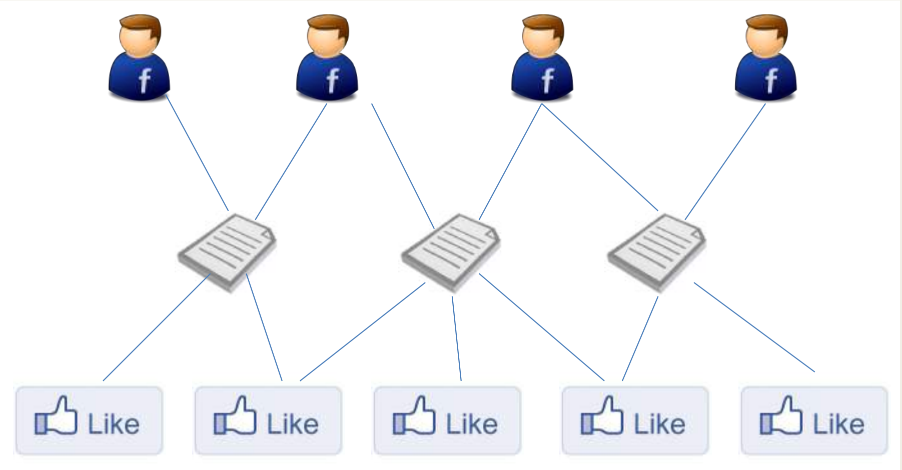
Test-retest
PCC

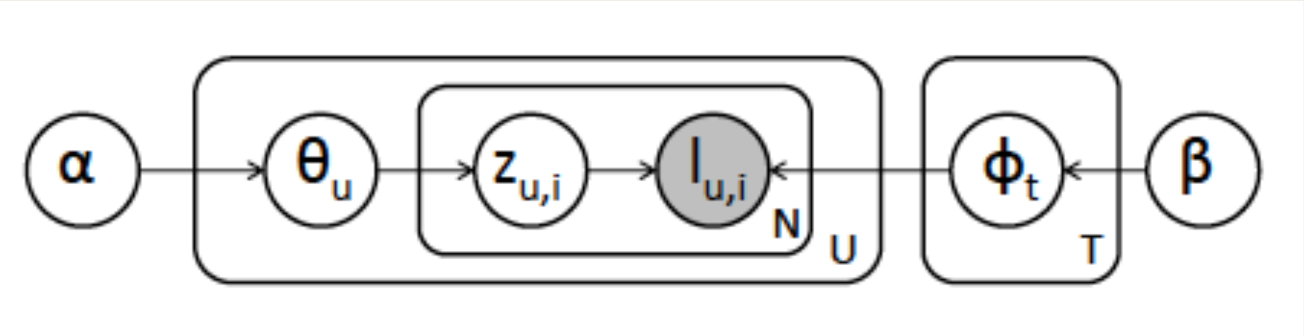
Prediction Accuracy [Kosinski et al., 2013]



More Features from LIKE: Topics

- A semantically coherent topic: a multinomial distribution of all LIKE items
- User is a mixture of a set of topics
  - Bag of words vs. bag of LIKEs





Further Improvement: Gender

[Kosinski-13] ⇨ SVD+LR

Model	Class	P	R	F	AUC	ACC
CAT+NB	Male	0.417	**0.885**	0.567	0.713	0.525
	Female	0.841	0.33	0.474	0.713	
CAT+LR	Male	0.757	0.547	0.635	0.726	0.779
	Female	0.787	0.905	0.842	0.726	
CAT+SVM	Male	0.721	0.678	0.699	0.835	0.795
	Female	0.831	0.858	0.844	0.835	
SVD+NB	Male	0.741	0.678	0.708	0.83	0.804
	Female	0.833	0.872	0.852	0.83	
SVD+LR	Male	0.823	0.747	0.783	0.852	0.865
	Female	0.87	0.913	0.891	0.842	
SVD+SVM	Male	0.831	0.789	0.809	0.863	0.874
	Female	**0.889**	0.913	0.901	0.863	
LDA+NB	Male	0.699	0.782	0.739	0.84	0.837
	Female	0.874	0.818	0.845	0.84	
LDA+LR	Male	0.884	0.731	0.8	0.873	0.872
	Female	0.867	**0.948**	0.906	0.873	
LDA+SVM	Male	**0.886**	0.781	**0.83**	**0.935**	**0.888**
	Female	0.888	0.945	**0.916**	**0.935**	

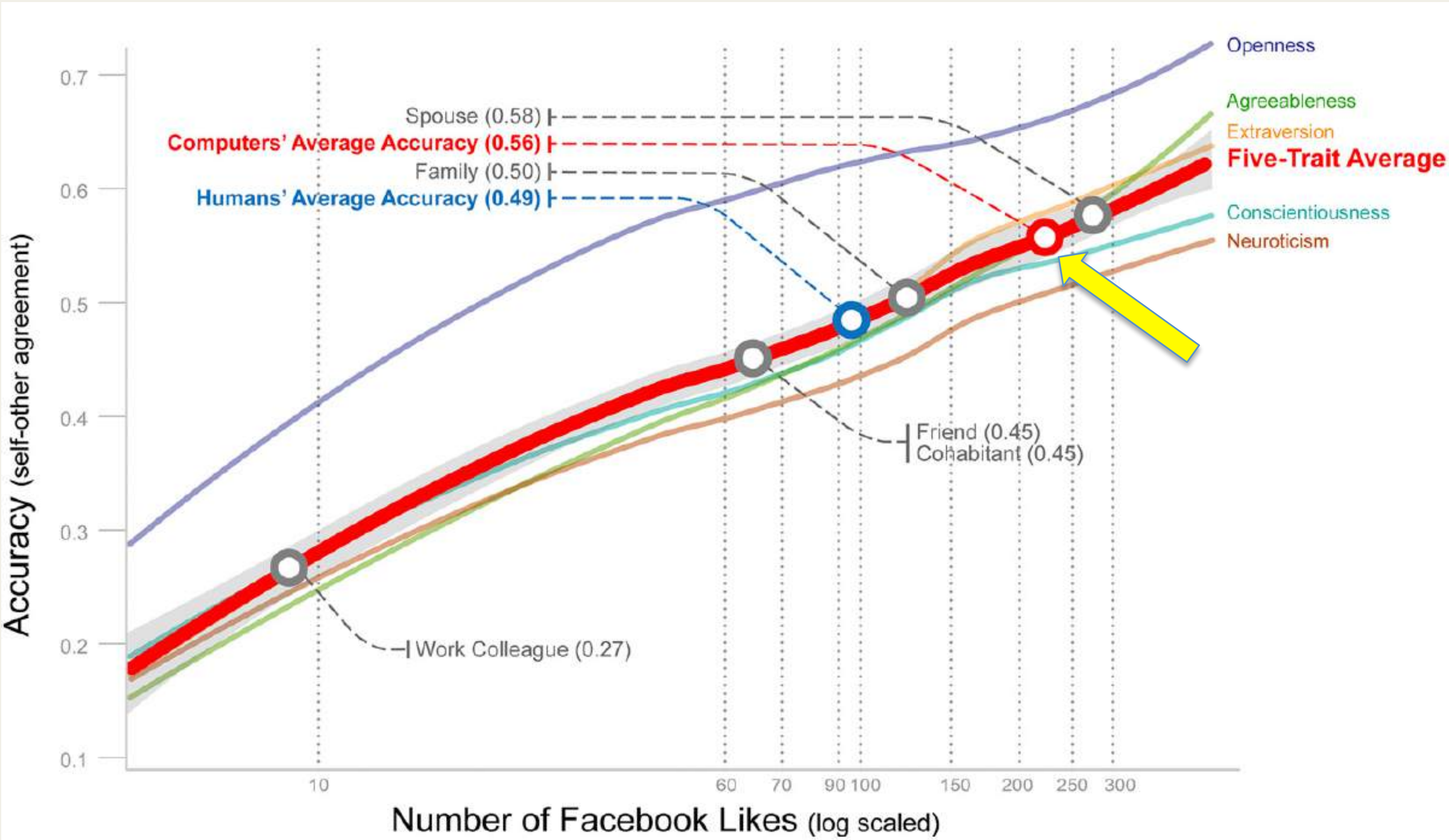
Further Improvement: Age

[Kosinski-13]

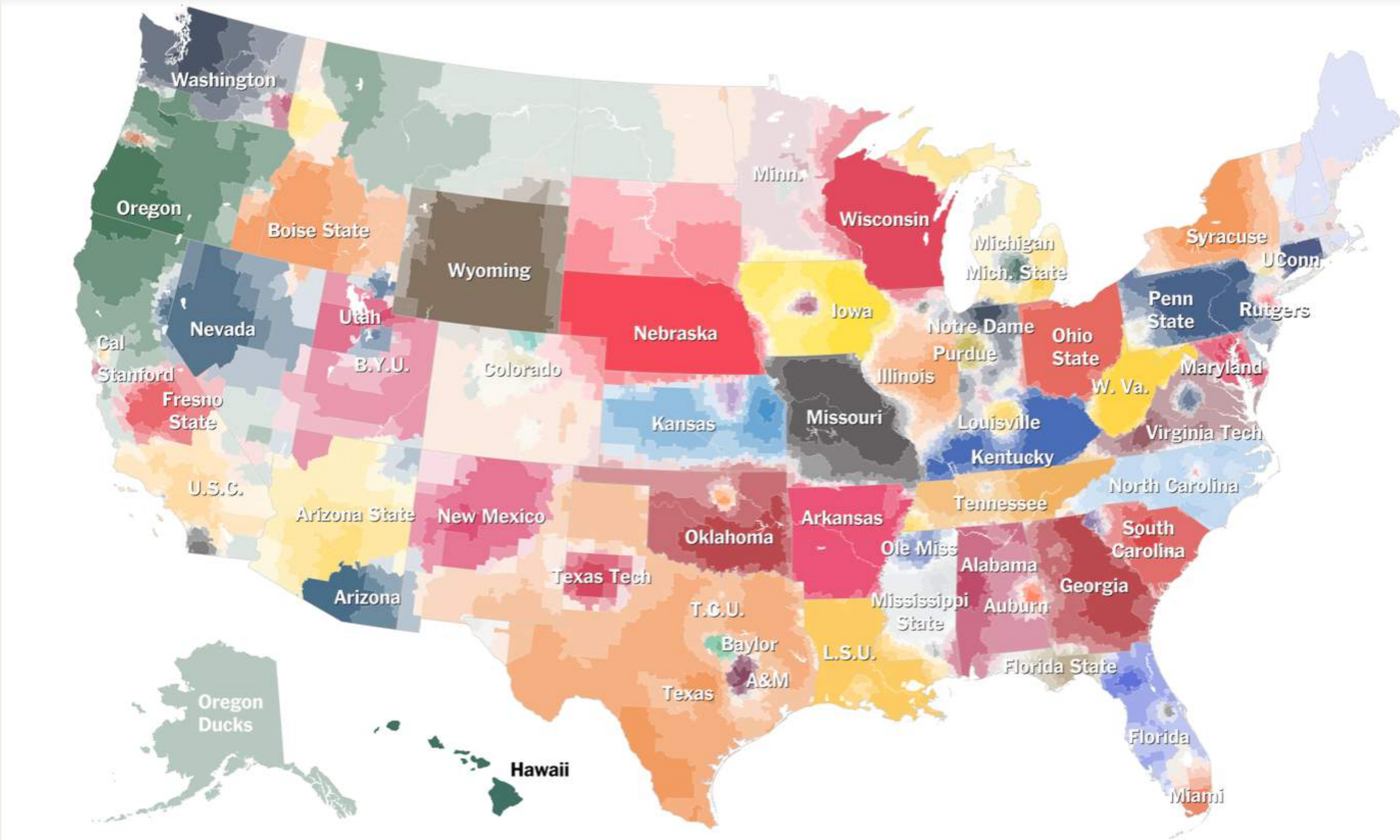
SVD+NB	20-	0.545	**0.626**	0.583	0.796	0.548
	20-30	0.655	0.547	0.596	0.685	
	30-40	0.363	**0.413**	0.386	**0.784**	
	40-50	0.265	0.29	0.277	0.756	
	50+	0.13	0.204	0.159	0.73	
SVD+LR	20-	0.659	0.539	0.593	**0.838**	0.583
	20-30	0.667	0.674	0.67	**0.757**	
	30-40	0.391	0.403	0.397	0.764	
	40-50	0.146	0.189	0.165	0.762	
	50+	0.167	0.273	0.207	0.765	
SVD+SVM	20-	**0.683**	0.59	0.633	0.821	0.643
	20-30	0.651	**0.806**	0.72	0.704	
	30-40	0.402	0.296	0.341	0.728	
	40-50	0.311	0.178	0.226	0.82	
	50+	0.4	0.122	0.188	0.778	
LDA+NB	20-	0.616	0.609	0.613	0.819	0.601
	20-30	**0.689**	0.608	0.646	0.719	
	30-40	0.313	0.374	0.341	0.739	
	40-50	0.226	**0.29**	0.254	0.794	
	50+	0.214	**0.367**	0.271	0.831	
LDA+LR	20-	0.619	0.552	0.584	0.792	0.623
	20-30	0.633	0.746	0.685	0.674	
	30-40	0.401	0.318	0.355	0.756	
	40-50	0.321	0.234	0.27	0.788	
	50+	0.304	0.143	0.194	0.758	
LDA+SVM	20-	0.677	0.624	**0.65**	0.826	**0.668**
	20-30	0.688	0.795	**0.738**	0.731	
	30-40	**0.453**	0.38	**0.413**	0.752	
	40-50	**0.4**	0.28	**0.33**	**0.831**	
	50+	**0.565**	0.265	**0.361**	**0.829**	

Predicting Personality using LIKE [Youyou et al., 2015]

- Human vs. machine in predicting personality

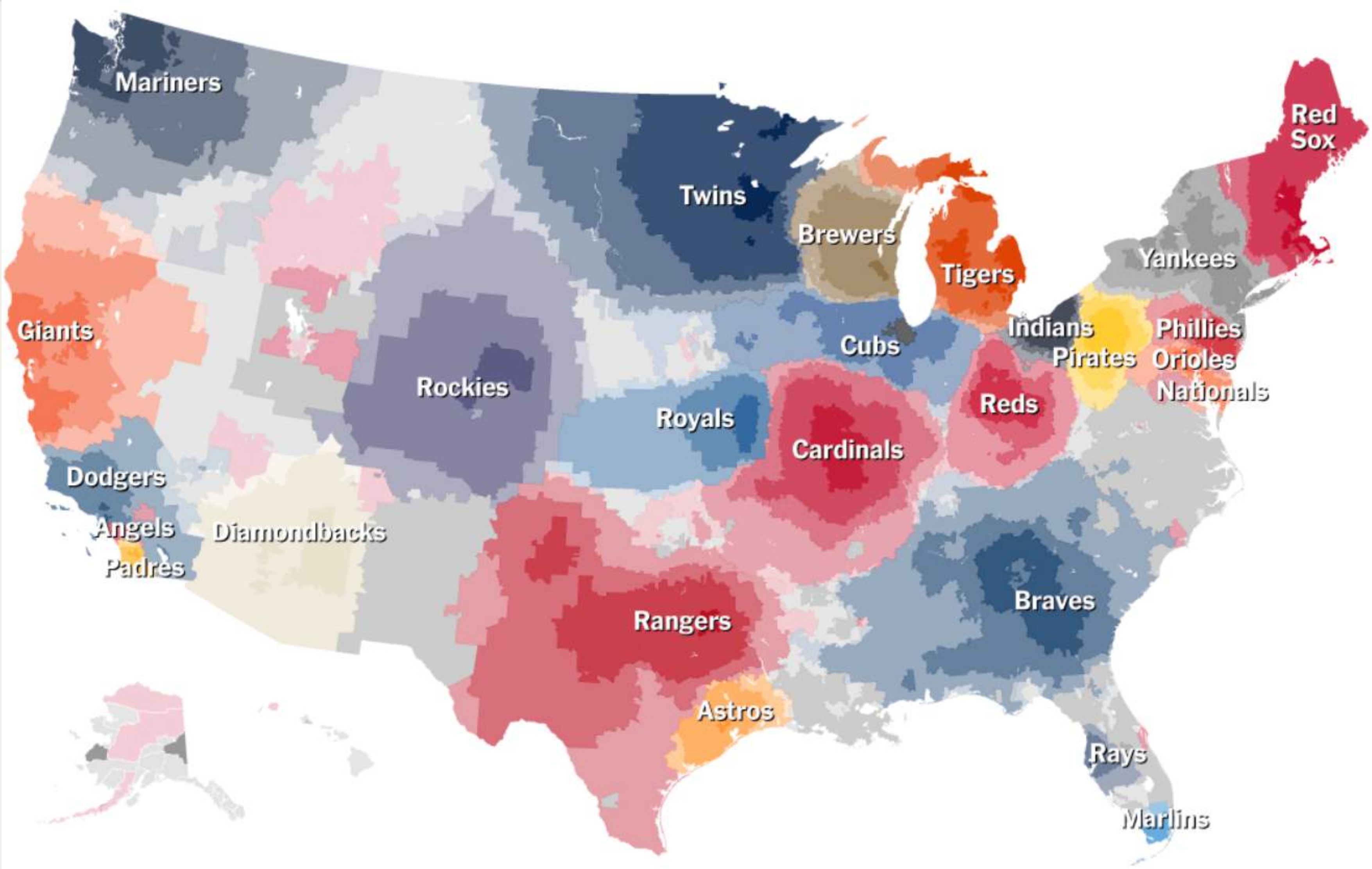


Football Fandom using LIKE



http://nyti.ms/1CIhun0

Baseball Fandom using LIKE



http://nyti.ms/1tF5e2W

Generation LIKE

- Pew report (2013)
  - 47% of all American teens and 82% of all American young adults own a smartphone
  - 81% of teens and 83% of young adults use social media
  - 93% of teens and young adults are online

- Q: Do teens use LIKE differently?

PBS Frontline, 2014

http://www.pbs.org/wgbh/pages/frontline/generation-like/

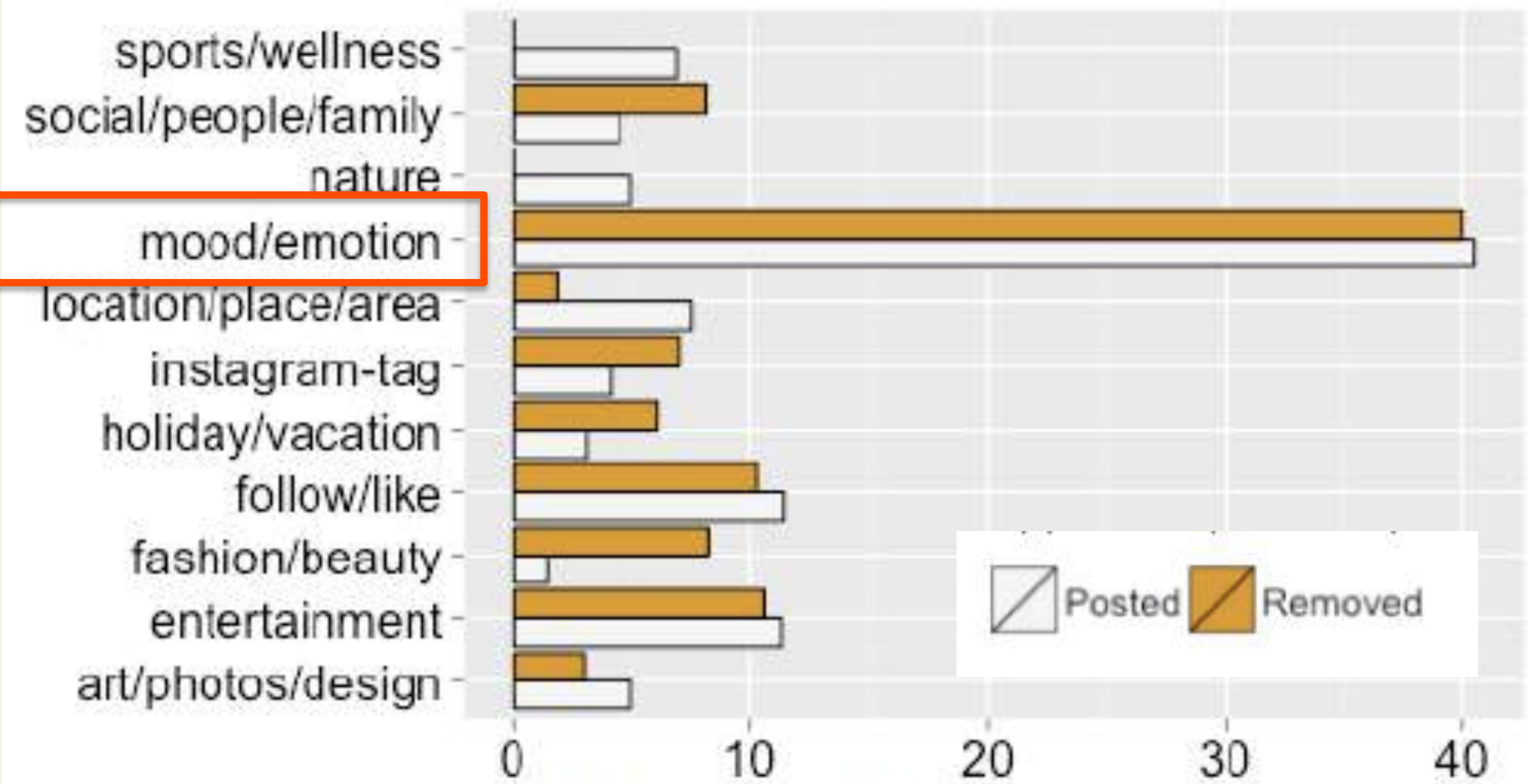
Teens vs. Adults w.r.t. LIKE [Han et al., TR]

	Teens (13,533)		Adults (13,352)	
	Median	SD	Median	SD
# Photos	110	272	175	487
# Likes	3,293	29,851	2,150	24,829
# Tags	446	2,595	294	2,511
# Comments	175	1,016	35	1,023
# Followers	401	3,683	348	5,700
# Followings	286	2,045	272	2,699

- Instagram study shows:
  - Teens post less photos but add more LIKEs

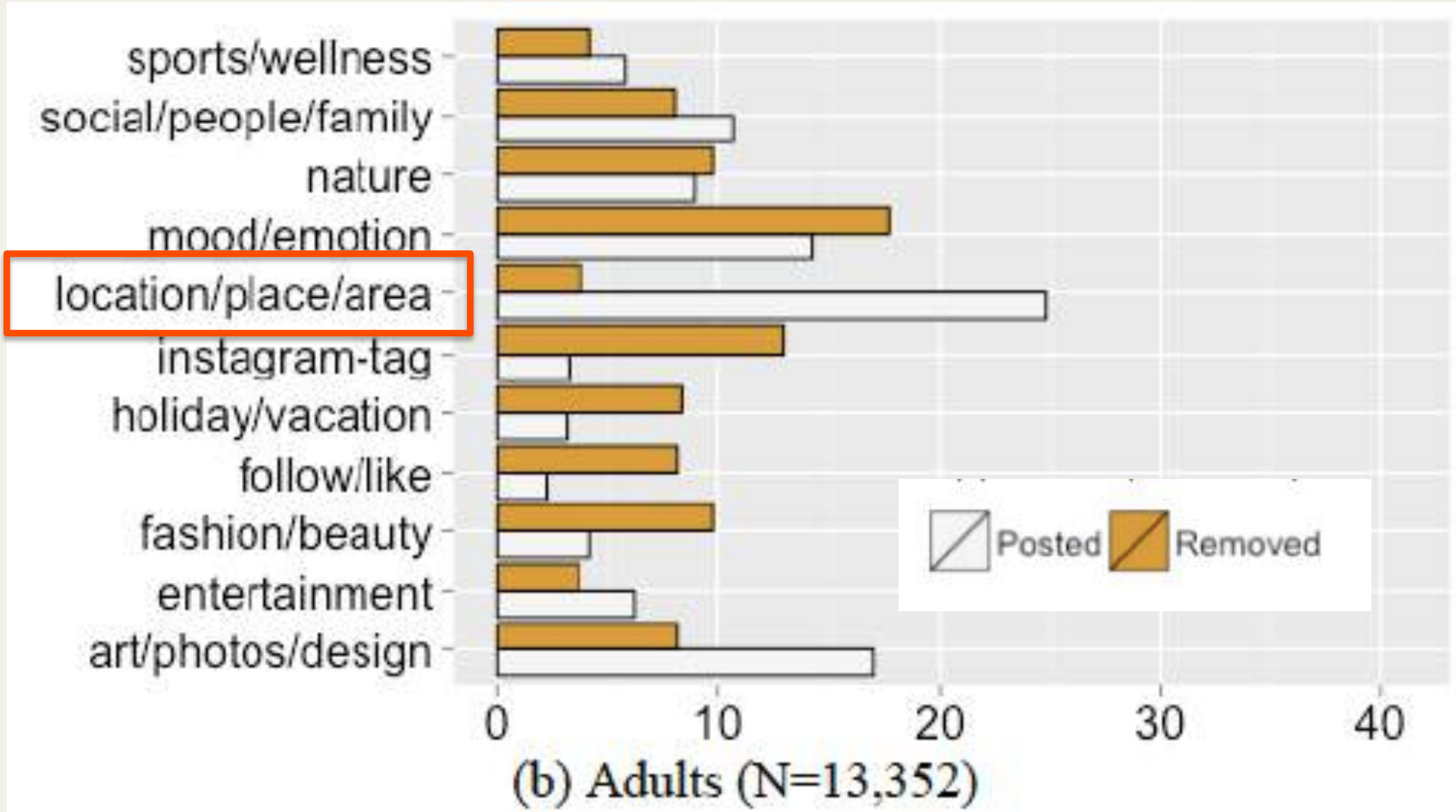
	Teens (13,553)	Adults (13,552)
# Likes / # Photos	56.36	35.31
# Tags / # Photos	6.42	4.70
# Comments / # Photos	2.52	1.06

Teens vs. Adults w.r.t. LIKE [Han et al., TR]

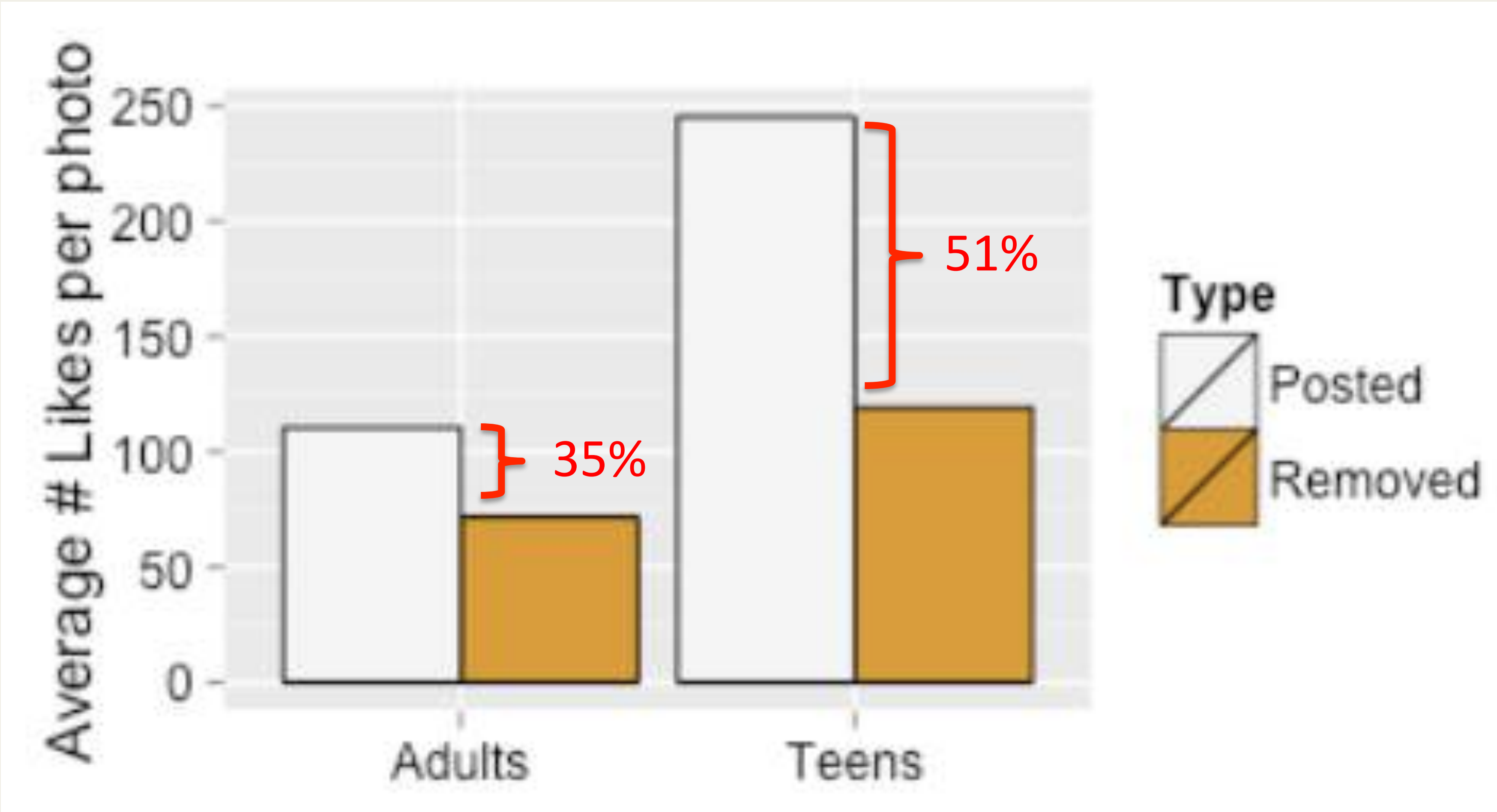


(a) Teens (N=13,533)

Teens vs. Adults w.r.t. LIKE [Han et al., TR]



Teens vs. Adults w.r.t. LIKE [Han et al., TR]



Teens vs. Adults w.r.t. LIKE [Han et al., TR]

- Two-tailed t-test, 5-point Likert scale
- Q: “Do you want to … ?”

	Teens	Adults
1. Receive more Likes*	3.59 (1.10)	3.20 (1.30)
2. Have more followers*	3.47 (1.23)	3.13 (1.34)
3. Look cool**	3.47 (1.24)	2.98 (1.34)
4. Increase visibility of photos+	3.33 (1.19)	3.00 (1.17)
5. Become more popular*	3.22 (1.20)	2.80 (1.28)
6. Frequency of usage	2.81 (1.40)	2.26 (1.31)

Note: $^{+}p < 0.10$; $^{*}p < 0.05$; $^{**}p < 0.01$

Teens vs. Adults w.r.t. LIKE [Han et al., TR]

- Two-tailed t-test, 5-point Likert scale
- Q: "In looking at other's photos, do you consider … ?"

	Teens	Adults
1. Matches my interest	4.19 (0.79)	4.08 (0.80)
2. Like its content	4.27 (0.78)	4.12 (0.77)
3. Like its quality	4.17 (0.78)	4.06 (0.84)
4. Poster has a lot of followers*	2.75 (1.32)	2.42 (1.31)
5. Photo has a lot of Likes*	2.86 (1.34)	2.42 (1.26)
6. Photo has a lot of Comments*	2.81 (1.37)	2.45 (1.23)

Note: * $p < 0.05$

Outline

- Introduction
- Understanding LIKEs
- Predicting LIKEs
- Aggregating LIKEs
- Summary

Motivation

- Predicting future # of LIKEs has commercial implications
  - Accurately
  - Early

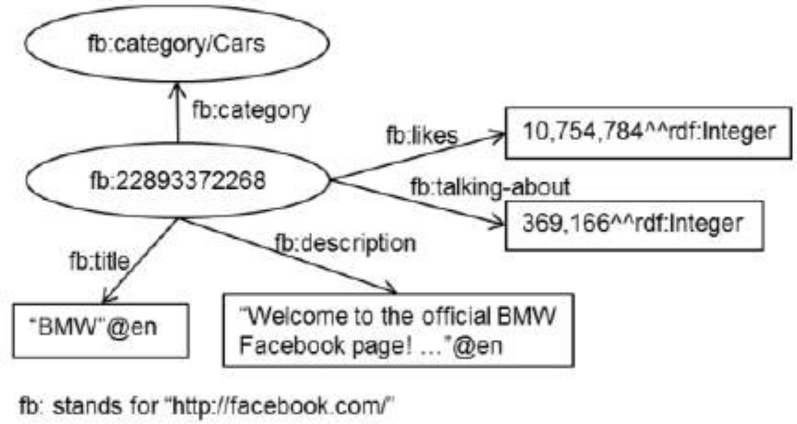
- Eg,
  - Viral video based marketing
  - Load balancing for popular videos

What Affects LIKE? [Jang et al., TR]

- Negative binomial regression (0.5M Instagram data)
  - Dependent variable : # of ***LIKEs***
  - IRR: Incident Rate Ratio

Variable	β	IRR	Std. err.	z	p
Followers	0.079	1.082	0.0004	173.0	< 0.0001
Photos	0.046	1.047	0.0004	102.3	< 0.0001
Comments	0.032	1.033	0.0002	114.5	< 0.0001
Tags	0.028	1.028	0.0002	120.7	< 0.0001
Follows	-0.005	0.994	0.0005	-9.7	< 0.0001

[Ohsawa and Matsuo, 2013]

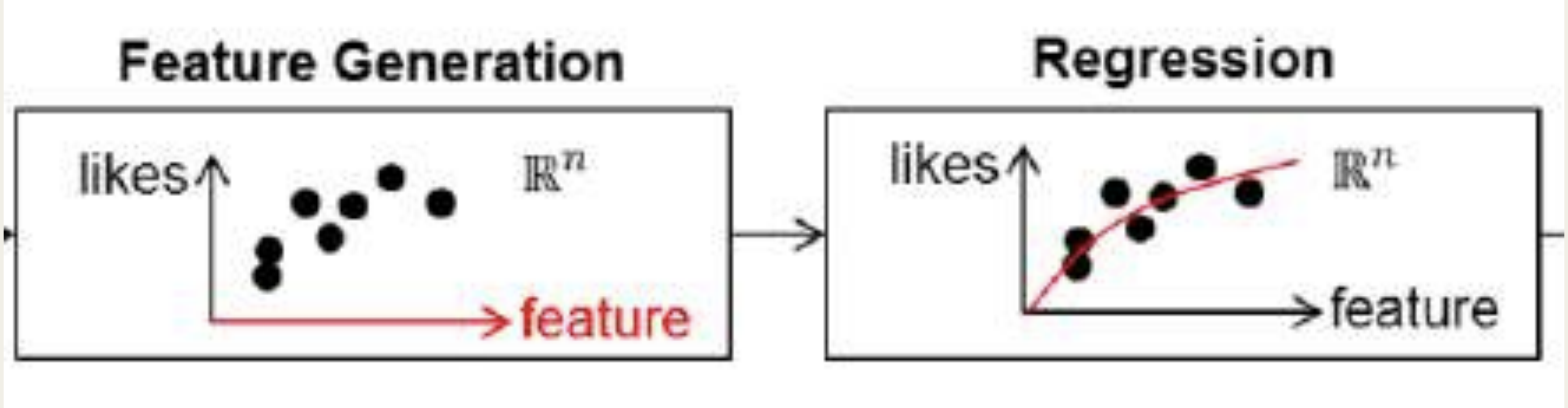


- Facebook LIKE dataset
  – 20M entities, 30B LIKEs
- DBPedia entities as dictionary
- Map Facebook entity to DBPedia entity

Name	Category	Likes
Facebook	Product	68,600,026
Rihanna	Musician	57,657,090
Lady Gaga	Musician	52,079,088
Harry Potter	Movie	52,069,702
Shakira	Musician	52,009,693
Michael Jackson	Musician	50,165,651
Family Guy	Tv show	46,445,463
Katy Perry	Musician	44,334,697
AKON	Musician	40,409,740
Music	Field of study	40,372,722

[Ohsawa and Matsuo, 2013]

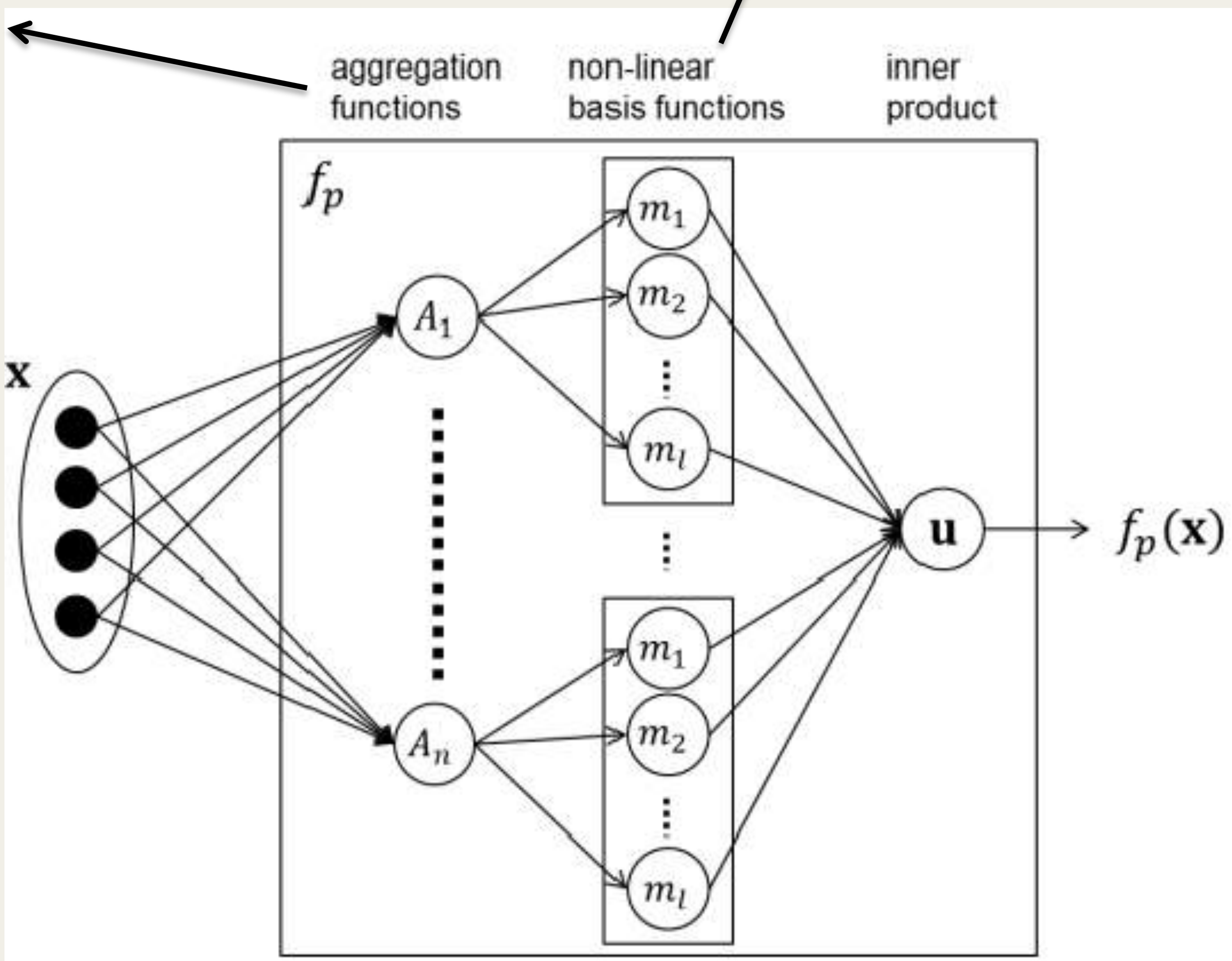
- Q: For each Facebook entity, predict # of LIKE
- Idea:
  - Link related entities first using RDF
  - Extract features from related entities
  - Solve the regression problem



[Ohsawa and Matsuo, 2013]

Eg, Logarithm, square root

Eg, Sum, Avg
Count, Max

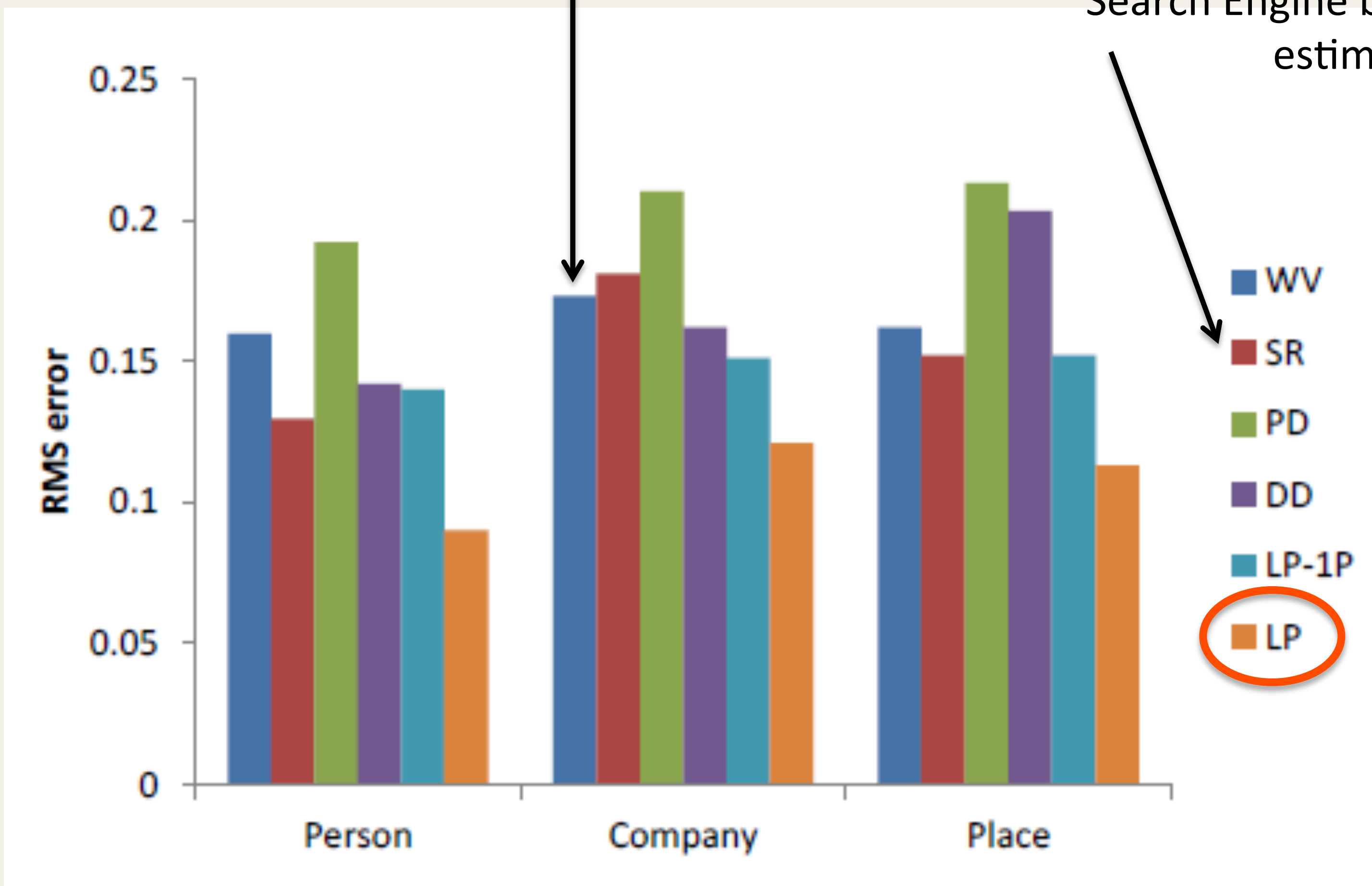


[Ohsawa and Matsuo, 2013]

- ~10% error rates

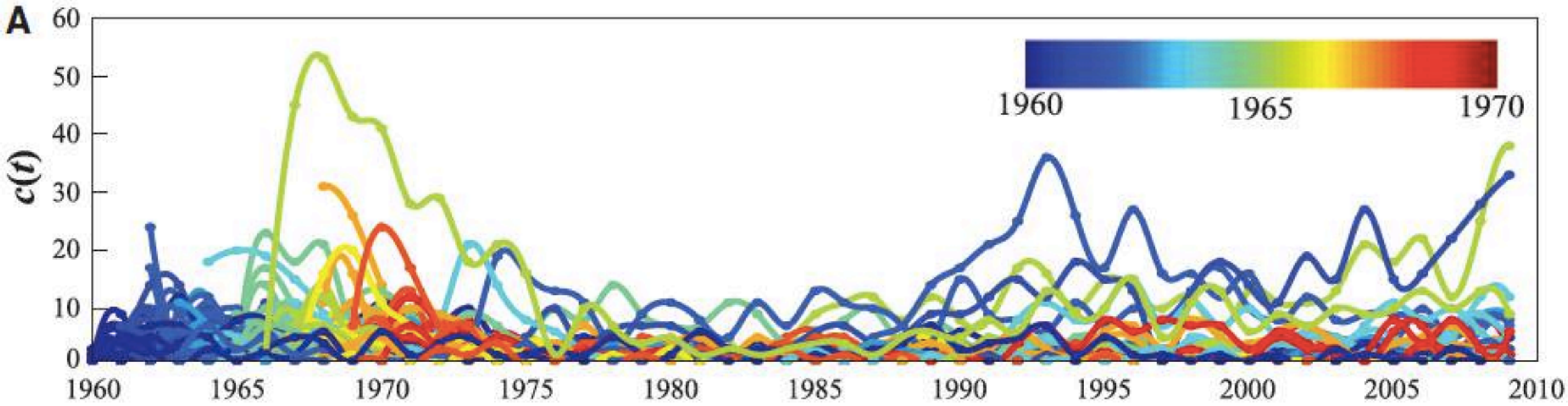
Wikipedia based estimation

Search Engine based estimation



Predicting Citation [Wang et al., 2013]

- Cell, PNAS, PRB 20 years citation dataset



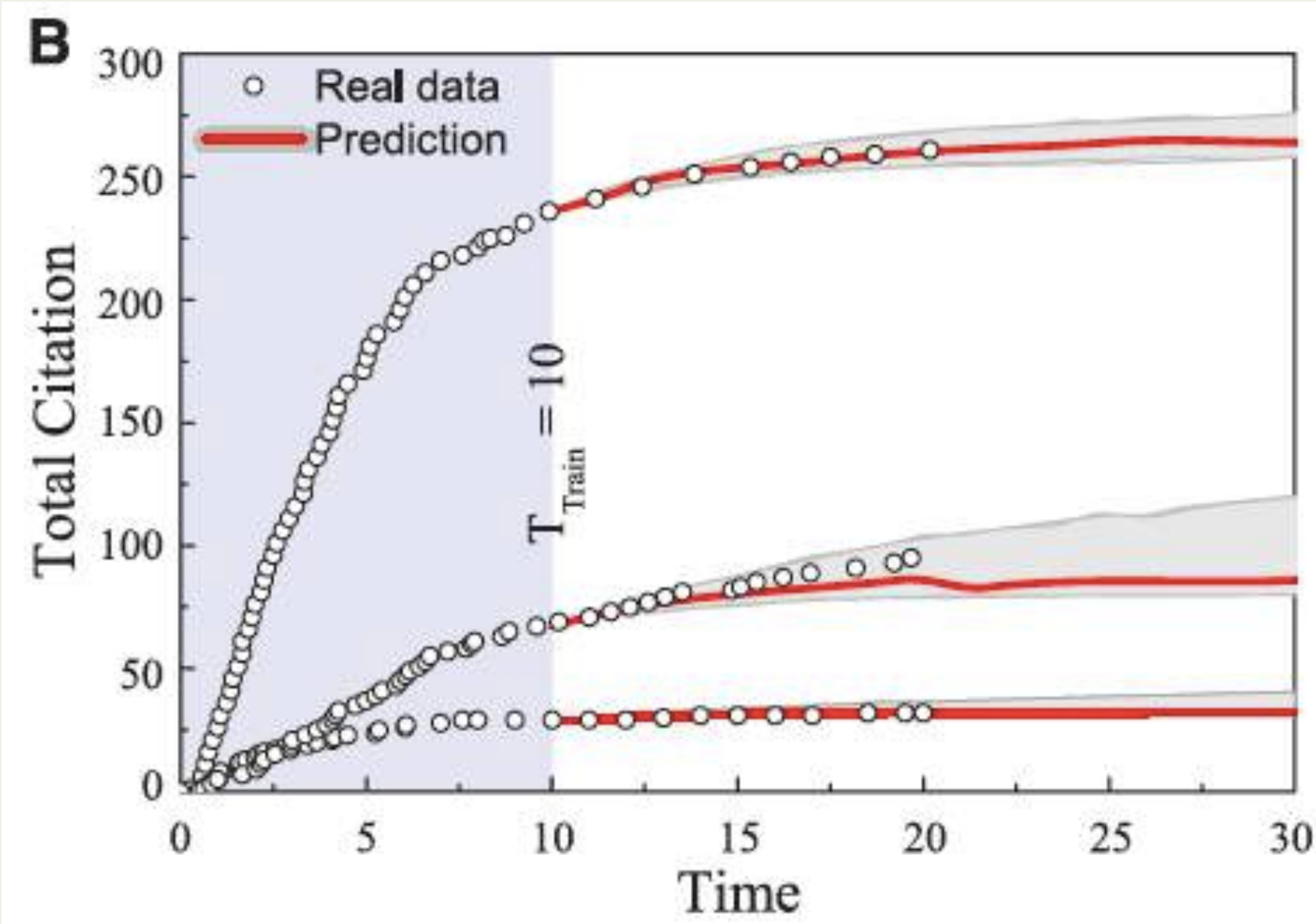
$$\Pi_i \sim \eta_i \, c_i^t \, P_i(t)$$

1) Preferential Attachment c_i^t

2) Aging $P_i(t)$

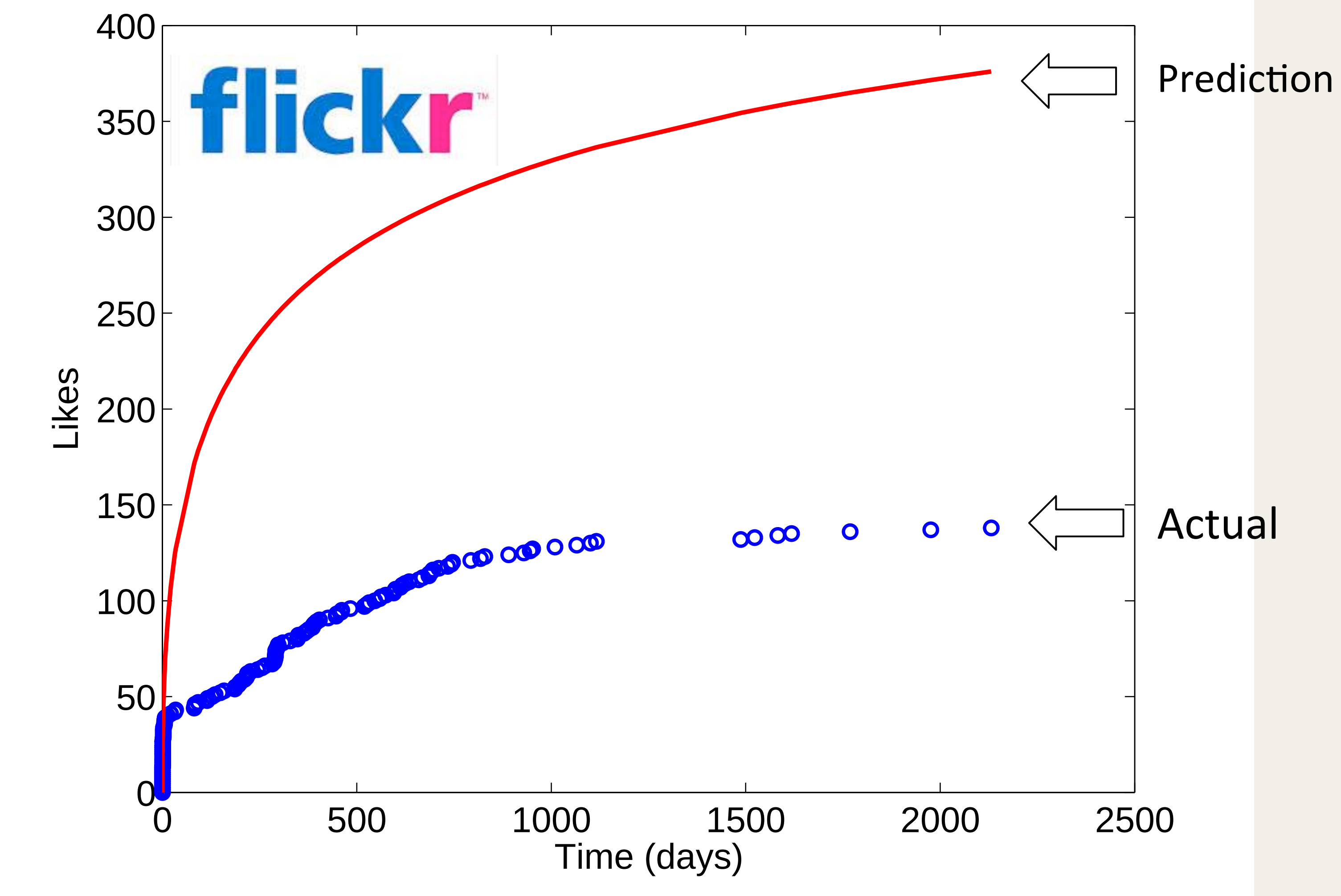
3) Intrinsic Novelty η_i

Predicting Citation [Wang et al., 2013]

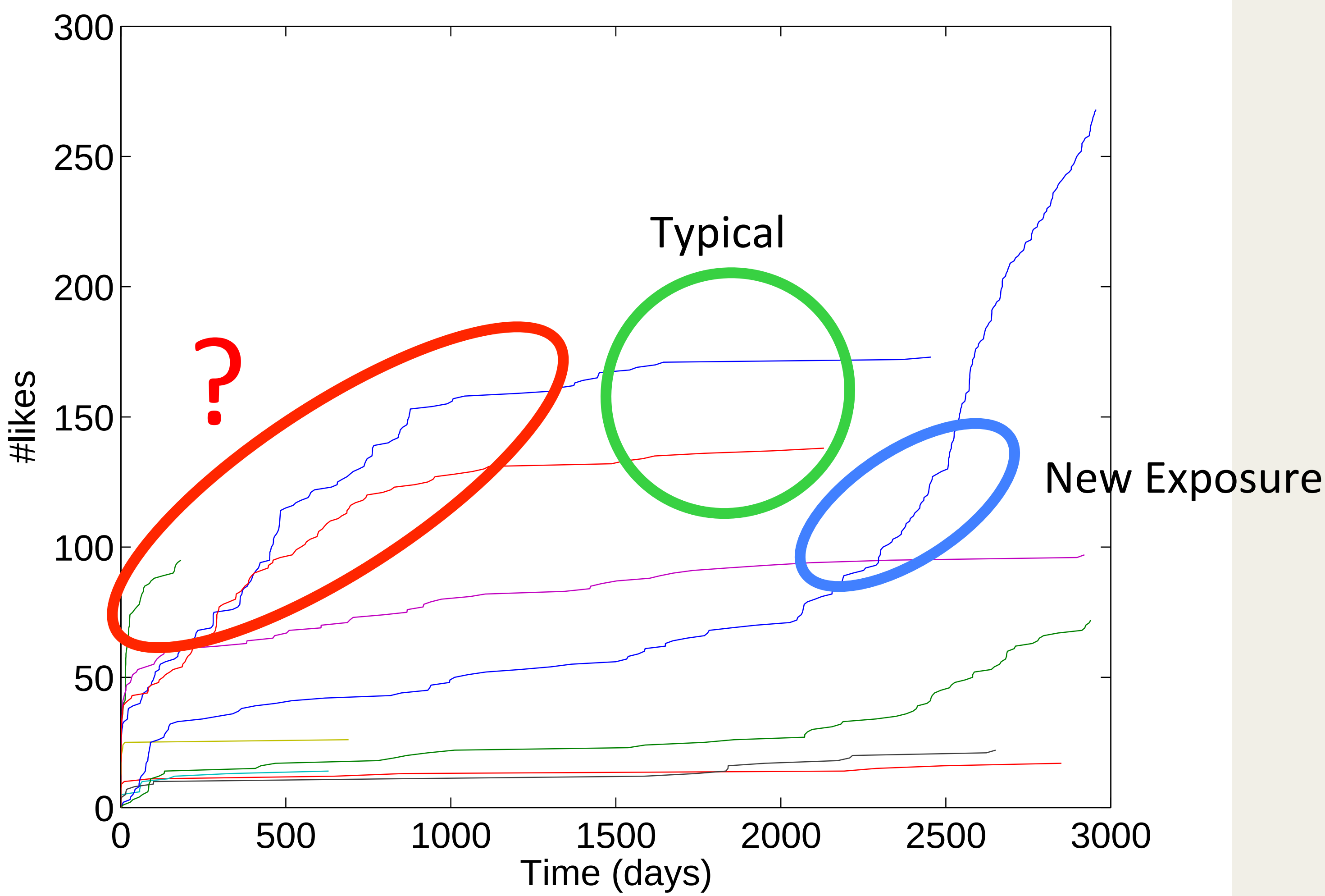


Predicting LIKE

- 6 years of Flickr dataset



Predicting LIKE

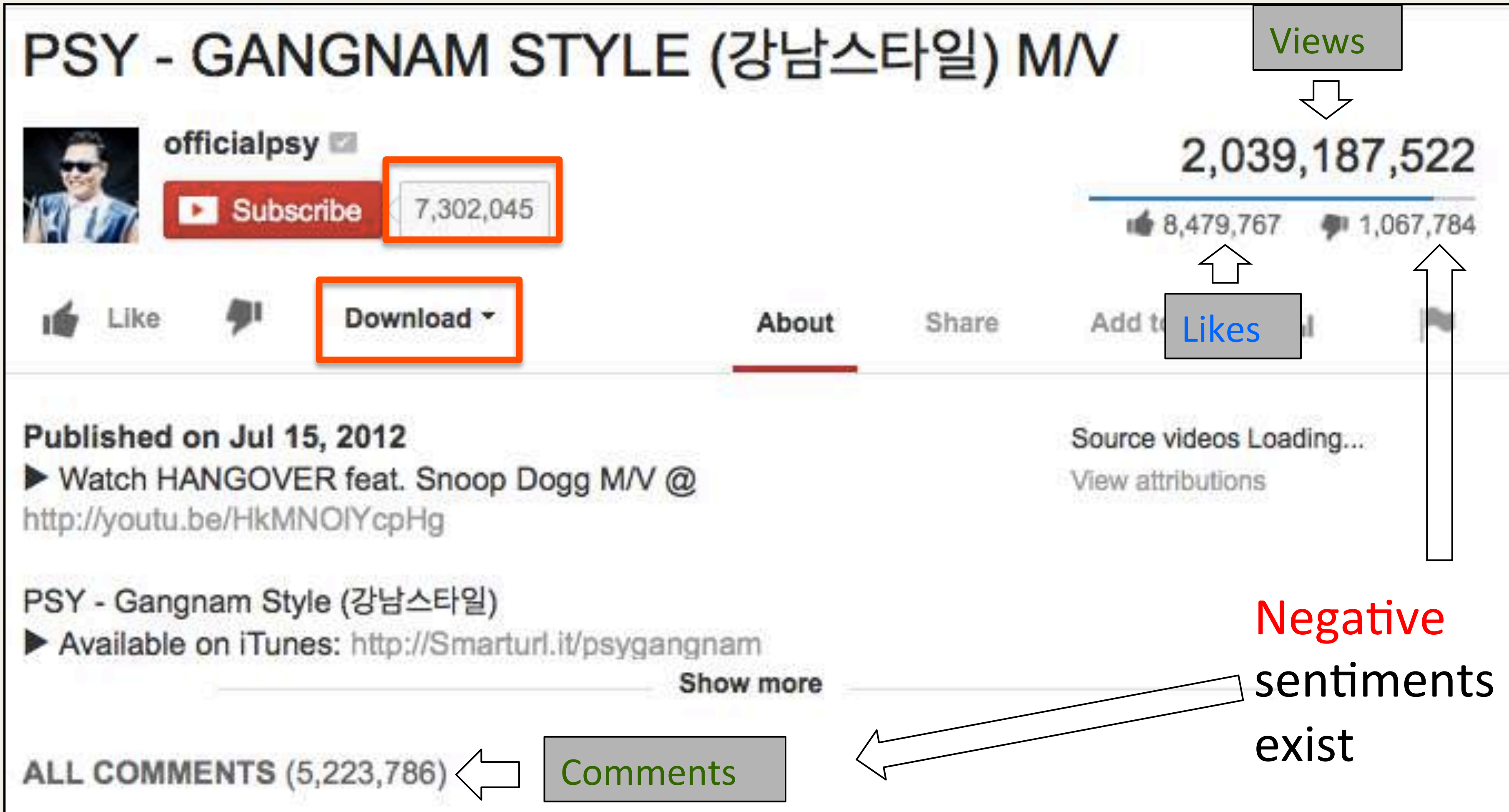


Heterogeneity of LIKE

- Eg, In YouTube
  1. A user viewed a clip
  2. A user downloaded a clip
  3. A user liked a clip
  4. A user commented to a clip
- Eg, In Twitter
  1. A user favorited a tweet
  2. A user re-tweets

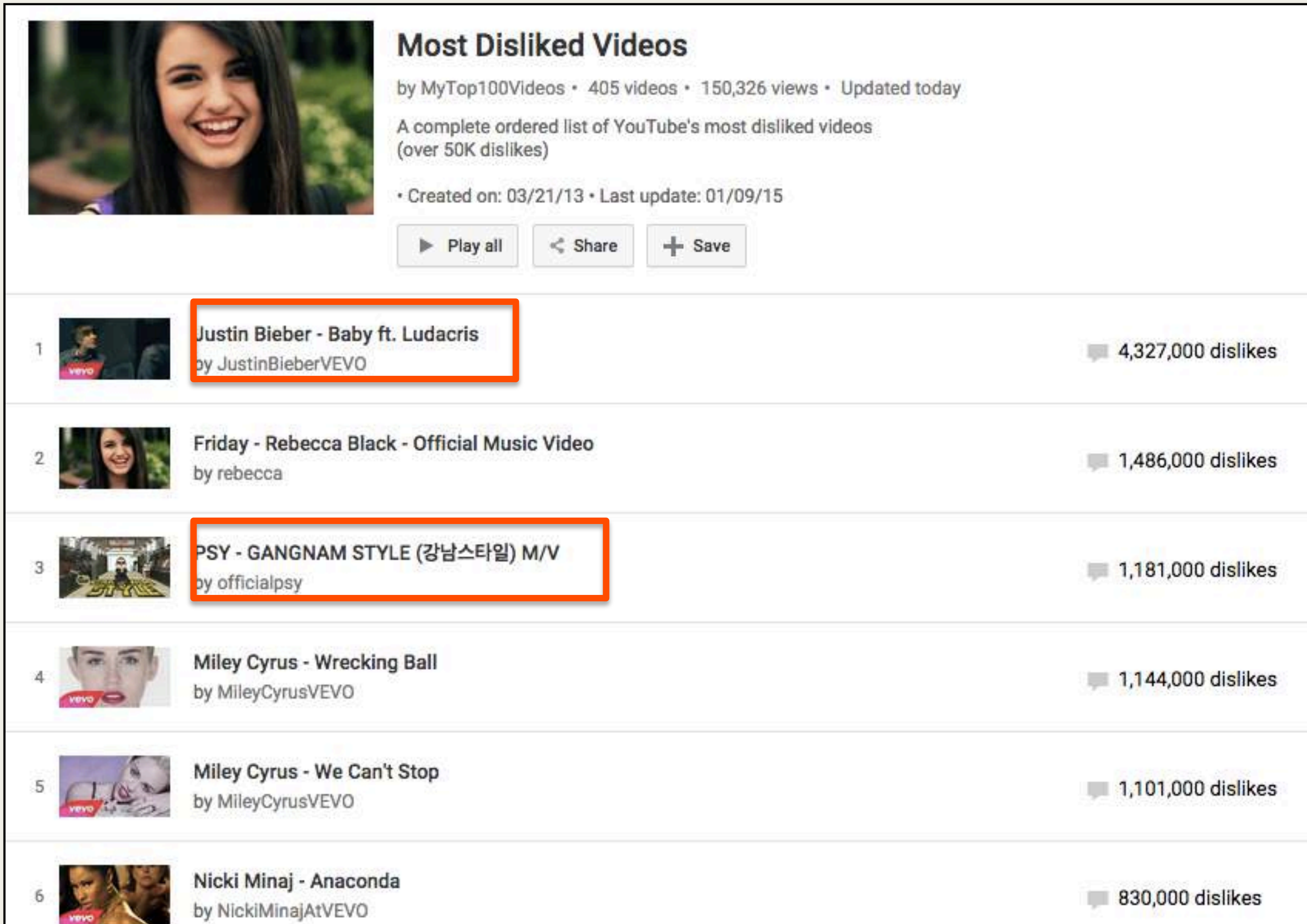
All some forms of Preference?

Heterogeneity of LIKE



View vs. LIKE vs. Dislike

View vs. LIKE vs. Dislike

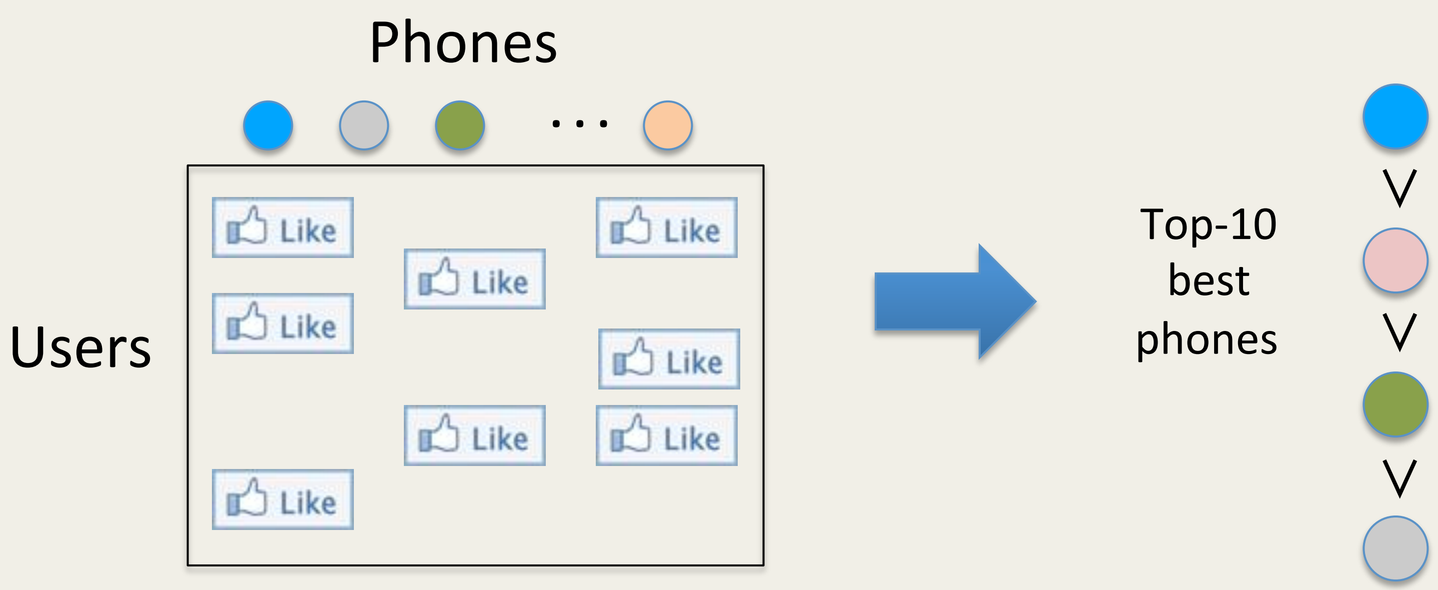


Outline

- Introduction
- Understanding LIKEs
- Predicting LIKEs
- Aggregating LIKEs
- Summary

Motivation

- Given users' LIKEs toward items in social media, how to form an aggregated ranking?
- Eg, In Facebook, what's the top-10 best phones in 2015 based on users' LIKEs?

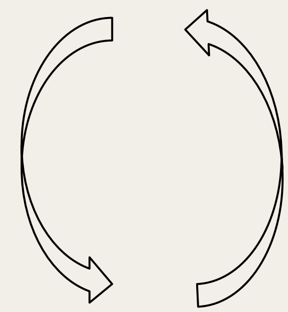


Modeling LIKEs

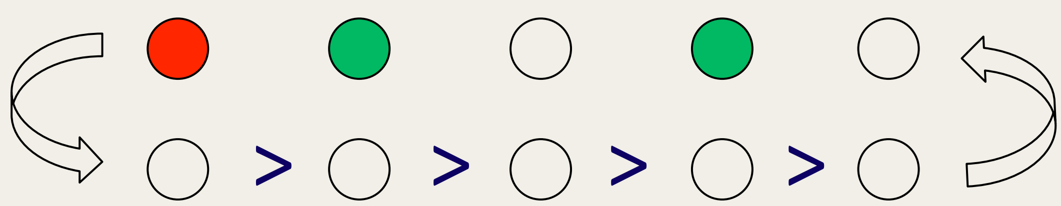
- *N*-ary LIKE

LIKE as Preference

- Quantitative
  – I rate “Godfather” 4 (out of 5)
- Qualitative
  – I prefer “Godfather” to “Pulp Fiction”



- Rating
- Ranking



Rating vs. Ranking as LIKE

- A rating of items assigns a numerical score to each item
  - [1 0 0 1 1]
  - When sorted, a rating of items form a ranked list
- A ranking of items is a ranked list of items
  - A ranking vector: a permutation of integers 1 .. N
  - [1 3 2 5 4]

Model 1: Rating Aggregation

- The **Rating Aggregation** Problem
  - Input: N rating lists: L_1, …, L_N
  - Output: An aggregated ranked list: L_A

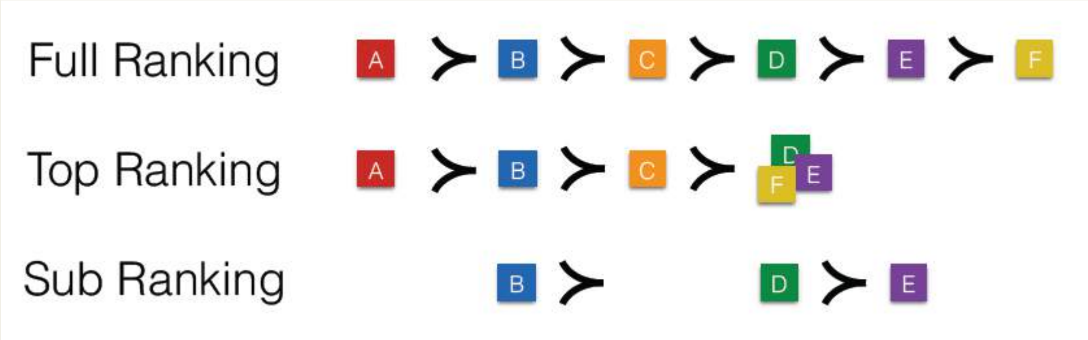
– Goal: Consensus at L_A

- Eg, NSF uses rating lists from panelists
  - P_1: 1VG, 3G
  - P_2: 2E, 2G
  - P_3: 1E, 1VG, 2G

$\Rightarrow$ $P_2 > P_3 > P_1$

Model 2: Rank Aggregation

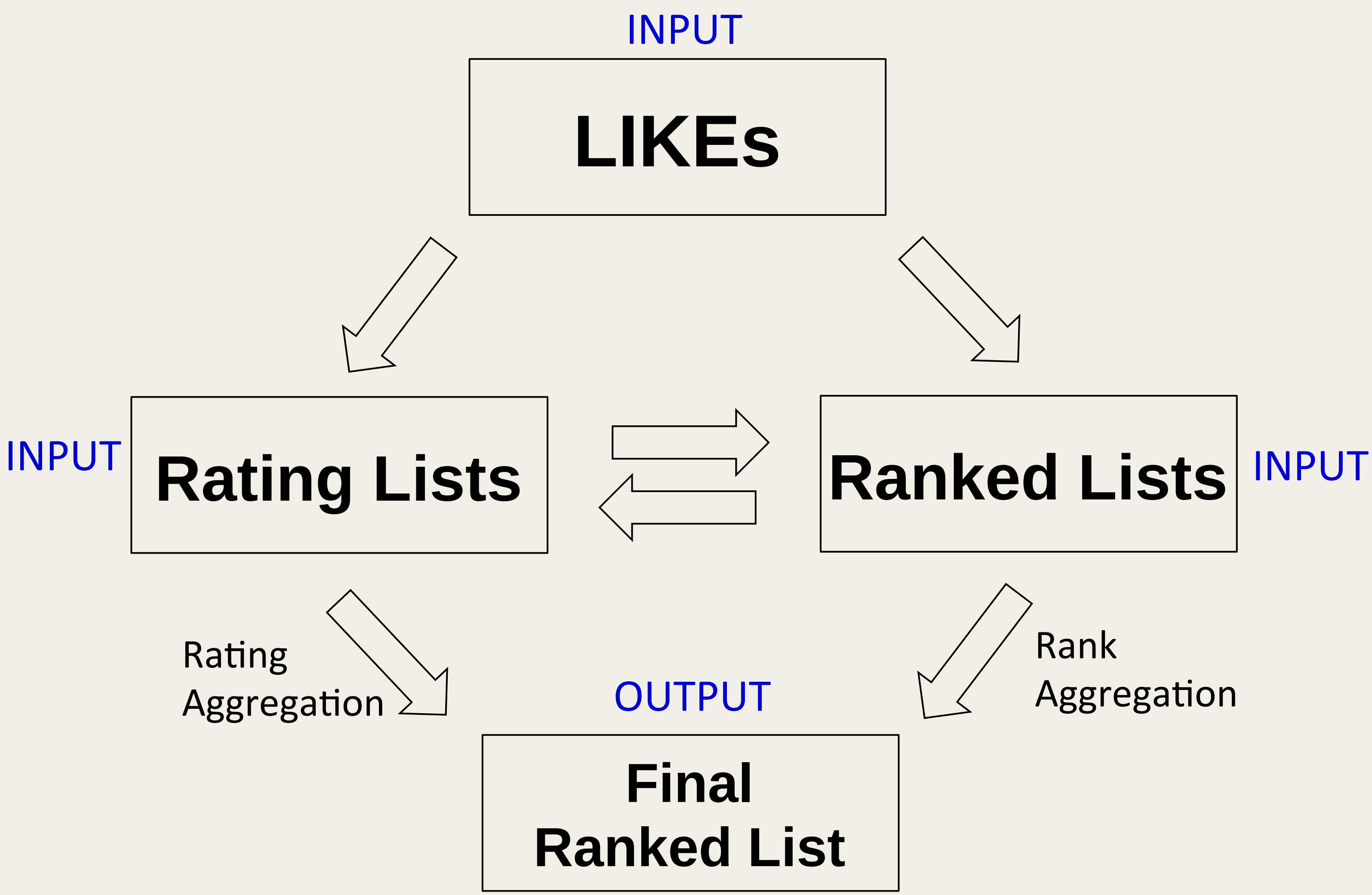
- The **Rank Aggregation** Problem
  - Input: N ranked lists: $L_1, \ldots, L_N$
  - Output: An aggregated ranked list: L_A



[Chen, 2014]

- Hypothesis
  - $Q(\text{worst}) \leq Q(L_i) \leq Q(L_A) \leq Q(\text{best})$
  - Q(): some imaginary quality function

Possible Workflow



1. LIKE (as Rating) Aggregation

- Solution #1
  - Convert ratings to rankings
  - Solve the rank aggregation problem
- Solution #2
  - Derive the final ranking from ratings directly

Eg, Movie Ranking

IMDb Charts

Top 250

As voted by regular IMDb users

Showing 250 Titles

Sort by: IMDb Rating

Rank & Title	IMDb Rating	Your Rating
1. The Shawshank Redemption (1994)	9.2	
2. The Godfather (1972)	9.2	
3. The Godfather: Part II (1974)	9.0	
4. The Dark Knight (2008)	8.9	
5. Pulp Fiction (1994)	8.9	
6. Schindler's List (1993)	8.9	
7. 12 Angry Men (1957)	8.9	
8. The Good, the Bad and the Ugly (1966)	8.9	

Eg, Doodle Scheduling

Consensus

Amazon Example

- 10 Amazon users rated 4 books in [1..5]
- Q: rank 4 books from highest to lowest?

	Book 1	Book 2	Book 3	Book 4
User 1	4	2	2	
User 2	3	1		2
User 3	1		2	
User 4	2	4	2	
User 5		3	2	5
User 6	2	3		
User 7		4	1	3
User 8	3	1	1	
User 9		3	2	5
User 10	2		2	4

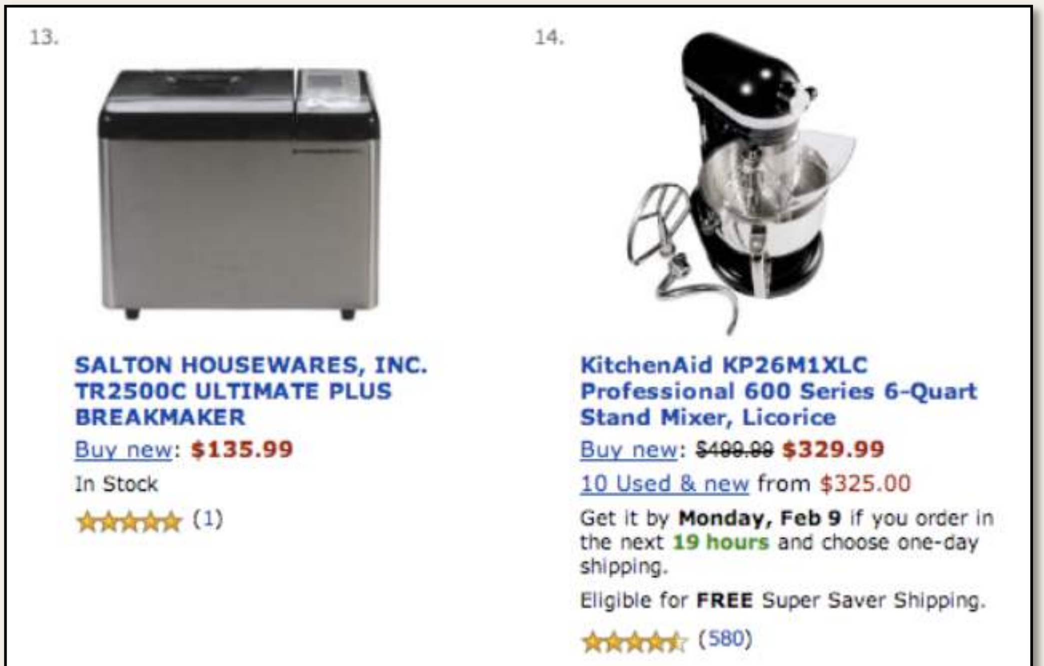
Average Rating Method

- Simple but could be counter-intuitive

	Book 1	Book 2	Book 3	Book 4
User 1	4	2	2	
User 2	3	1		2
User 3	1		2	
User 4	2	4	2	
User 5		3	2	5
User 6	2	3		
User 7		4	1	3
User 8	3	1	1	
User 9		3	2	5
User 10	2		2	4
AVG rating	2.43	2.63	1.75	3.8
Rank	3	2	4	1

Average Rating Method

- May not work well when rating lists have different lengths (ie, different # of ratings)



http://www.evanmiller.org/how-not-to-sort-by-average-rating.html

Centroid Rating Method [Langville and Meyer, 2012]

Theorem: for a set of ratings on n items, the best rating is given by the centroid vector $\boldsymbol{r}$:

$$r = \frac{Ke}{n}$$

where $\boldsymbol{K}$ is the skew-symmetric matrix of average rating differences and can be approximated by the score difference matrix $\boldsymbol{S}$, such that $k_{ij} = s_{ij} - s_{ji}$, where s_{ij} = (1) average rating given to item i by the user who rated both items i and j, and (2) 0 otherwise

Centroid Rating Method [Langville and Meyer, 2012]

	Book 1	Book 2	Book 3	Book 4
User 1	4	2	2	
User 2	3	1		2
User 3	1		2	
User 4	2	4	2	
User 5		3	2	5
User 6	2	3		
User 7		4	1	3
User 8	3	1	1	
User 9		3	2	5
User 10	2		2	4

$s_{12} = (4 + 3 + 2 + 2 + 3)/5 = 14/5$

S

	Book 1	Book 2	Book 3	Book 4
Book 1	0	14/5	12/5	5/2
Book 2	11/5	0	17/6	11/4
Book 3	9/5	10/6	0	7/4
Book 4	6/2	15/4	17/4	0

$k_{ij} = s_{ij} - s_{ji}$

K

	Book 1	Book 2	Book 3	Book 4
Book 1	0	3/5	3/5	-1/2
Book 2	-3/5	0	7/6	-1
Book 3	-3/5	-7/6	0	-10/4
Book 4	1/2	1	10/4	0

$k_{12} = s_{12} - s_{21} = 14/5 – 11/5 = 3/5$

$$r = \frac{Ke}{4}$$

	Centroid Vector *r*	Rank
Book 1	7/40 = 0.175	2
Book 2	-13/120 = -0.108	3
Book 3	-16/15 = -1.07	4
Book 4	1	1

2. LIKE (as Rank) Aggregation

- Old problem
  - Voting theory (social choice)
- Many disciplines
  - Economics
  - Political Science
  - Mathematics
  - Statistics
  - Computer Science
- Many modern applications

Eg, US College Football Ranking

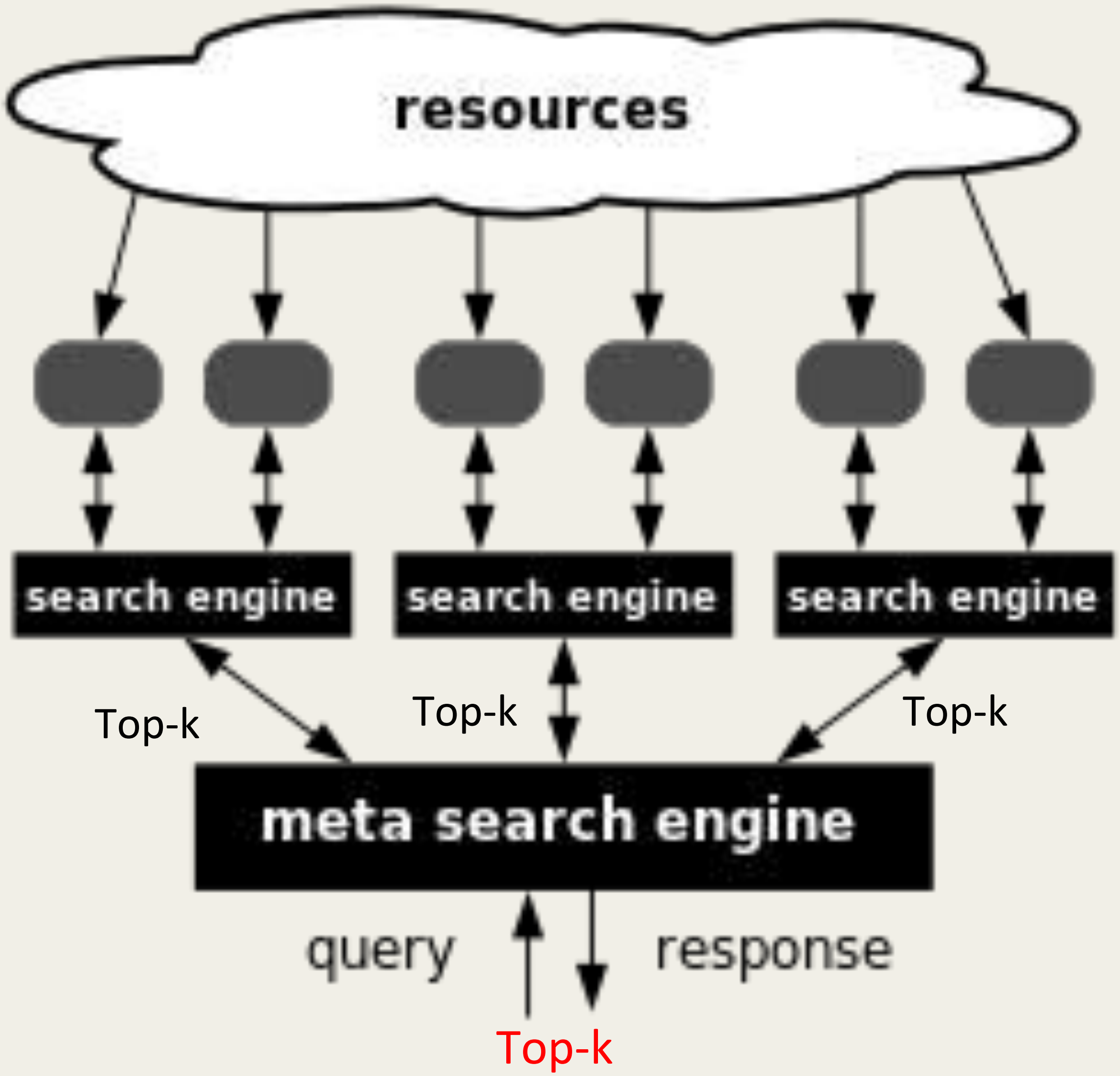
USAT CFCC	Coffey
Payne	Bihl
Laz Index	Donchess Inference
Kirkpatrick	England
RoundTable	AccuRatings

UCC	PAY	LAZ	KPK	RT	COF	BIH	DII	ENG	ACU	Rank, Team, Conf, Record			
1	1	1	1	1	1	1	1	1	1	1	Ohio St	B10	14-1
2	2	2	2	2	2	2	2	4	2	2	Oregon	P12	13-2
3	4	3	4	3	4	3	4	2	3	3	Alabama	SEC	12-2
4	3	4	3	4	3	5	3	3	4	4	TCU	B12	12-1
6	6	6	5	5	7	6	5	6	11	5	Michigan St	B10	11-2
7	7	7	6	7	6	11	8	7	8	6	Georgia	SEC	10-3
5	5	5	7	6	8	4	6	5	5	7	Florida St	ACC	13-1
8	8	9	9	10	5	7	7	10	7	8	Baylor	B12	11-2
9	11	8	13	11	11	12	9	14	9	9	Georgia Tech	ACC	11-3
13	10	13	11	8	9	10	11	9	10	10	Mississippi	SEC	9-4

UCC	PAY	LAZ	KPK	RT	COF	BIH	DII	ENG	ACU	Rank, Team, Conf, Record			
10	9	12	10	9	10	8	10	8	6	11	UCLA	P12	10-3
12	12	14	14	14	12	9	13	12	14	12	Mississippi St	SEC	10-3
11	18	10	8	20	15	13	15	11	12	13	Missouri	SEC	11-3
15	17	11	12	15	20	17	14	13	17	14	Wisconsin	B10	11-3
14	13	22	17	17	17	15	12	17	19	15	Clemson	ACC	10-3
18	19	20	16	18	16	18	20	15	15	16	Auburn	SEC	8-5
16	15	15	21	13	14	19	17	22	22	17	Boise St	MWC	12-2
20	20	19	15	19	19	22	18	16	16	18	USC	P12	9-4
17	21	18	18	21	18	16	19	19	13	19	Arizona St	P12	10-3
21	22	21	19	22	21	21	22	18	20	20	Kansas St	B12	9-4

http://www.masseyratings.com/cf/compare.htm

Eg, Meta Search Engine



Average Rank Method

- Integers representing rank are averaged
- Simple but ties are frequent

	Voter 1 rank	Voter 2 rank	Voter 3 rank	Avg	Avg Rank
Brazil	1	3	2	2	**1**
Argentina	2	1	3	2	1
Germany	3	2	1	2	1
Netherlands	4	5	4	4.3	4
Colombia	5	4	5	4.7	5

Average Rank Method

- Counter-intuitive if input ranks have different lengths

	Voter 1 rank	Voter 2 rank	Voter 3 rank	…	Voter 10 rank	Avg	Avg Rank
Brazil	3	2	1	1	1	1.2	**2**
Argentina	4	3	2	2	2	2.3	**4**
Germany	5	4	3			4	**5**
Netherlands	2	1				1.5	**3**
Colombia	1					1	**1**

Average Rank Method

- Counter-intuitive if input ranks have a skewed distribution

	Voter 1 rank	Voter 2 rank	Voter 3 rank	Avg	Avg Rank
Brazil	1	1	5	2.3	**2**
Argentina	2	3	3	2.6	**3**
Germany	3	2	1	2	**1**
Netherlands	4	5	4	4.3	**5**
Colombia	5	4	2	3.6	**4**

Median Rank Method [Fagin et al., 2003]

- Median of each rank is used for aggregated rank
- Used in Olympic figure skating

	Voter 1 rank	Voter 2 rank	Voter 3 rank	Median	Median Rank
Brazil	1	3	4	3	**2**
Argentina	2	1	3	2	**1**
Germany	3	4	1	3	**2**
Netherlands	4	2	5	4	**4**
Colombia	5	5	2	5	**5**

Median Rank Method [Fagin et al., 2003]

- Could address some of counter-intuitive problems of average rank method

	Voter 1 rank	Voter 2 rank	Voter 3 rank	…	Voter 10 rank	Median	Median Rank
Brazil	3	2	1	1	1	1	1
Argentina	4	3	2	2	2	2	3
Germany	5	4	3			3	5
Netherlands	2	1				2	3
Colombia	1					1	1

Borda Count Method

- By Jean-Charles de Borda in 1770
- For each ranked list, candidate gets points = # of outranked candidates + 1
  - – *n* for 1st preference, *n-1* for 2nd, … *1* for the last
- Final ranking is based on the sum of points

	Voter 1 rank	Voter 2 rank	Voter 3 rank	Borda Count	Borda Rank
Brazil	1	1	3	13	1
Argentina	2	3	2	11	3
Germany	3	2	1	12	2
Netherlands	4	4	4	6	4
Colombia	5	5	5	3	5

Rank	Point
1	5
2	4
3	3
4	2
5	1

Borda Count Method

- Good for consensus winner
- Not necessarily good for majority winner
- Eg,
  - 6/10 voted for Brazil as 1st → majority winner
  - But, Germany is the Borda winner

	6 voters	2 voters	2 voters
1st	Brazil	Germany	Argentina
2nd	Germany	Argentina	Germany
3rd	Argentina	Netherlands	Brazil
4th	Netherlands	Colombia	Colombia
5th	Colombia	Brazil	Netherlands

⇒

Country	Borda Point	Borda Rank
Brazil	6x5 + 2x1 + 2x3 = 38	2
Germany	6x4 + 2x5 + 2x4 = 42	1
Argentina	6x3 + 2x4 + 2x5 = 36	3
Netherlands	6x2 + 2x3 + 2x1 = 20	4
Colombia	6x1 + 2x2 + 2x2 = 10	5

Condorcet Method

- By Marquis de Condorcet in 1785
- Candidate wins by majority rule against each other candidate in one-on-one contests
  - Condorcet Winner: candidate who defeats every other candidate in pairwise majority rule election

	6 voters	2 voters	2 voters
1st	Brazil	Germany	Argentina
2nd	Germany	Argentina	Germany
3rd	Argentina	Brazil	Brazil

→

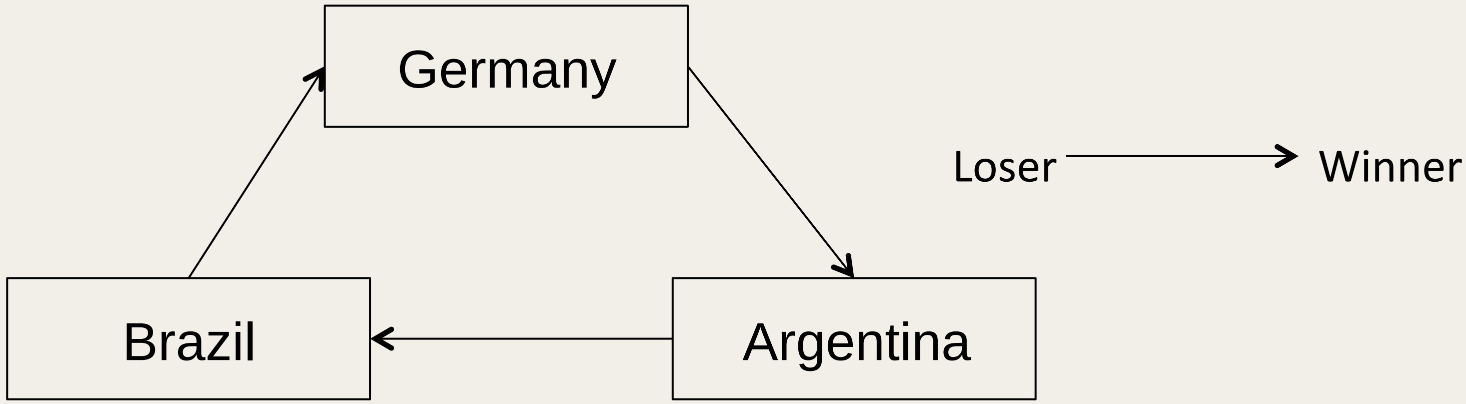
	Brazil	Germany	Argentina
Brazil		B>G: 6, B<G: 4	B>A: 6 B<A: 4
Germany			G>A: 10
Argentina			

Brazil > Germany: 6 & Brazil > Argentina: 6

→ Brazil is Condorcet winner

Condorcet Paradox

- Voter 1: Germany > Brazil > Argentina
- Voter 2: Brazil > Argentina > Germany
- Voter 3: Argentina > Germany > Brazil



- Condorcet winner may not exist

Borda winner ≠ Condorcet winner

- Borda Count
  - Brazil: 6x3 + 4x1 = 22
  - Germany: 6x2 + 4x3 = 24 ⇦ Winner
  - Argentina: 6x1 + 4x2 = 14
- Condorcet Method
  - Brazil beats both Germany and Argentina ⇦ Winner

	6 voters	4 voters
1st	Brazil	Germany
2nd	Germany	Argentina
3rd	Argentina	Brazil

⇨

	Brazil	Germany	Argentina
Brazil		B>G: 6, B<G: 4	B>A: 6 B<A: 4
Germany			G>A: 10
Argentina			

Copeland Method

- By A. H. Copeland in 1951
- Ranked by:

  # of pairwise wins – # of pairwise losses

- Good fit for sports: round-robin tournaments

	6 voters	2 voters	2 voters
1st	Brazil	Germany	Argentina
2nd	Germany	Argentina	Germany
3rd	Argentina	Brazil	Brazil
4th	Colombia	Colombia	Colombia

	Brazil	Germany	Argentina	Colombia
Brazil		B>G: 6, B<G: 4	B>A: 6 B<A: 4	B>C: 10
Germany			G>A: 10	G>C: 10
Argentina				A>C: 10
Colombia				

	Wins	Losses	Win - Loss	Copeland Rank
Brazil	3	0	3	1
Germany	2	1	2	2
Argentina	1	2	-1	3
Colombia	0	3	-3	4

Uses "upvotes – downvotes" as part of story ranking algorithm

http://amix.dk/blog/post/19588

Toward Perfect Voting Theory

- Arrow's Impossibility Theorem, 1951
  - Nobel prize in Economics in 1972
- NO "ranked" voting systems with ≥ 3 candidates can simultaneously satisfy 4 properties:
  - Every voter is able to rank candidates in any order → Unrestricted domain
  - Ranking higher should not hurt a candidate → Monotonicity
  - No single voter should have disproportionate control over an election → Non-dictatorship
  - If *A* is preferred to *B* out of the choice set {*A*,*B*}, it must remain the same even if expanded to {*A*,*B*,*C*}, → Independence of Irrelevant Alternatives (IIA)

IIA Criteria Controversy

- IIA: relative rankings within subsets should be maintained when expanded to supersets
- Eg, US 2000 Presidential Election
  - In Florida, Gore was Condorcet winner
    - Gore > Nader & Gore (+Nader) > Bush (+Buchanan)
  - But Bush won in plurality scheme
    - Bush > Gore > Nader
  - Violation of IIA
    - {Gore, Bush} → Gore > Bush
    - {Gore, Bush, Nader} → Bush > Gore

Which is a better method? [Kumar 2008]

- Kemeny's Proposal, 1959
  - Ideal rank aggregation method yields the aggregated ranking that is the least distant from input rankings
- How to measure distance between rankings?
  - Kendall Tau distance
  - Spearman's Footrule distance
  - Lots more

Kendall Tau Distance

- Degree to which one rank (dis)agrees with another
  - Ranges from -1 to 1
- Eg,
  - p: Argentina > Brazil > Colombia > Germany
  - q: Brazil > Germany > Argentina > Colombia
    - Argentina-Brazil : X
    - Argentina-Colombia : O
    - Argentina-Germany : X
    - Brazil-Colombia : O
    - Brazil-Germany : O
    - Colombia-Germany : X

(# agreements - # disagreements) / all #
= (3 – 3) / 6 = 0

Spearman's Footrule Distance

- L1 distance between two ranks p and q
  - |p-q|

- Eg,
  - p: Argentina > Brazil > Colombia > Germany
  - q: Brazil > Germany > Argentina > Colombia
    - Argentina: |1 – 3| = 2
    - Brazil = |2 – 1| = 1
    - Colombia = |3 – 4| = 1
    - Germany = |4 – 2| = 2

(2 + 1 + 1 + 2)
= 6

Spearman's Weighted Footrule Distance

- Weighting to reflect that disagreements near top are more troubling than those near bottom
  - $|p-q|/\min(p,q)$
- Eg,
  - p: Argentina > Brazil > Colombia > Germany
  - q: Brazil > Germany > Argentina > Colombia
    - Argentina: $|1 – 3|/1 = 2$
    - Brazil = $|2 – 1|/1 = 1$
    - Colombia = $|3 – 4|/3 = 1/3 = 0.3$
    - Germany = $|4 – 2|/2 = 2/2 = 1$

$(2 + 1 + 0.3 + 1)$
$= 4.3$

Findings [Kumar 2008]

- Diaconis-Graham inequality shows:
  - Kendal distance ≤ Footrule distance ≤ 2 x Kendal distance

- Optimal Aggregation: Given a set of rankings R, an optimal aggregation A is the one that has the minimum sum of Kemeny distances to all rankings in R

Findings [Kumar 2008]

- Kemeny optimal aggregation problem is NP-hard
  - NP-hard even for 4 rankings
- Given an optimal aggregation A, C-approximate aggregation A' satisfies:
  - Sum of distances to A' $\leq$ C x sum of distances to A
  - At worst, C times off from the optimal solution

Findings [Kumar 2008]

- Copeland rank aggregation is a 6-approximation to Kemeny optimal aggregation
- Borda rank aggregation is a 5-approximation to Kemeny optimal aggregation
- Median rank aggregation is a 3-approximation to Kemeny optimal aggregation

Outline

- Introduction
- Understanding LIKEs
- Predicting LIKEs
- Aggregating LIKEs
- Summary

Summary

- LIKE as an interesting and novel lens to understand people and their lives in social media
- LIKE activities are correlated with personal traits to some extent
- Teens have somewhat different LIKE activities from other age groups
- Predicting # of LIKE accurately is still challenging
- Using LIKEs for global recommendation is doable
  ➔ More on recommendation in Part 2

References

- **[Bakhshi et al., 2014]** S. Bakhshi, D. Shamma, E. Gilbert, *Faces Engage Us: Photos with Faces Attract More Likes and Comments on Instagram*, ACM CHI, 2014
- **[Chen, 2014]** W. Chen**,** *How to Order Sushi: A Nonparametric Approach to Modeling Rank Data***,** Harvard University, MS Thesis, 2014
- **[Fagin et al., 2003]** R. Fagin, R. Kumar and D. Sivakumar. *Efficient similarity search and classification via rank aggregation*, ACM SIGMOD 2003
- **[Ferrara et al, 2014]** E. Ferrara, R. Interdonato, A. Tagarelli, *Online Popularity and Topical Interests through the Lens of Instagram*, ACM Hypertext, 2014
- **[Gleich and Lim, 2011]** D. Gleich and L.-H. Lim, *Rank Aggregation via Nuclear Norm Minimization*, ACM KDD, 2011
- **[Han et al., 2015]** K. Han**,** J. Jang, D. Lee, *Exploring Tag-based Like Networks*, ACM CHI, Extended Abstracts, 2015
- **[Han et al., TR]** K. Han, J. Jang, H. Jia, P. Shih, D. Lee, *All About "Likes"? Comparing Teens' and Adults' Behaviors in Instagram,* Penn State Univ., Technical Report, 2015
- **[Jang et al., 2015]** J. Jang, K. Han, P. Shih, D. Lee, *Generation LIKE: Comparative Characteristics in Instagram*, ACM CHI, 2015

References

- **[Jang et al., TR]** J. Jang, K. Han, D. Lee, *Comprehensive Analysis on Like Activities in Instagram*, Penn State Univ., Technical Report, 2015
- **[Kosinski et al., 2013]** M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, Proc. Natl. Acad. Sci., 2013
- **[Kumar, 2008]** R. Kumar**,** *Rank Aggregation*, U. Rome Seminar, 2008
- **[Langville and Meyer, 2012]** A. Langville, C. Meyer, *Who's #1? The Science of Rating and Ranking*, Princeton University Press, 2012
- **[Ohsawa and Matsuo, 2013]** S. Ohsawa and Y. Matsuo. *Like prediction: modeling like counts by bridging facebook pages with linked data*, WWW, 2013
- **[Robbins, 2013]** I. Robbins, What is the Meaning of 'Like'?: The First Amendment Implications of Social-Media Expression, Federal Courts Law Review, 2013
- **[Rossi, 2011]** F. Rossi, *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*, *Synthesis* Lectures on Artificial Intelligence and Machine Learning, 5(4):1, 2011
- **[Wang et al., 2013]** D. Wang et al., *Quantifying Long-term Scientific Impact*, Science, 2013
- **[Youyou et al., 2015]** W. Youyou, M. Kosinski, D. Stillwell, *Computer-based personality judgments are more accurate than those made by humans*, Proc. Natl. Acad. Sci., 2015

Part 2: Recommendation in Social Media

Huan Liu

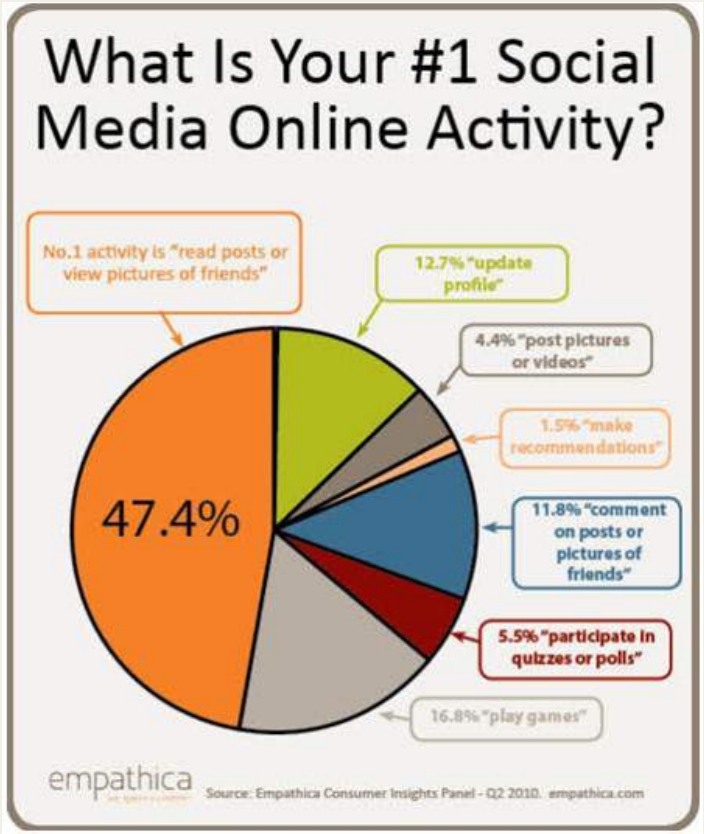
1

Outline

- Introduction
- Content Recommendation
- Location Recommendation
- Future Work

Social Media [Zafarani et al., 2014]

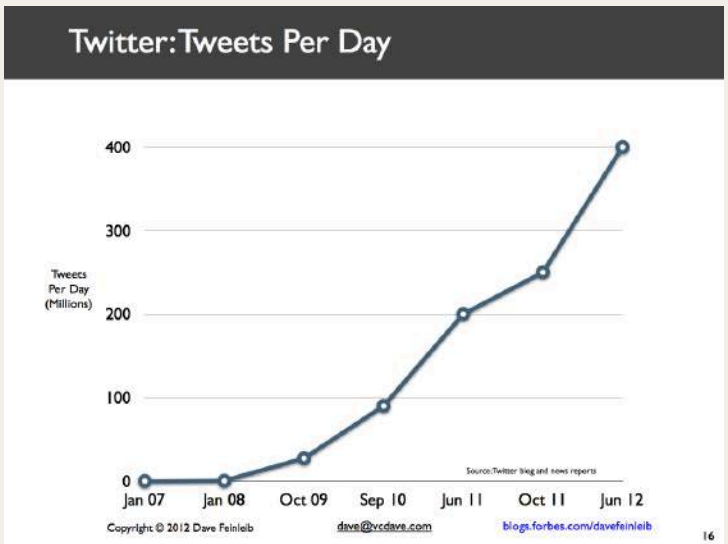
- Social media greatly enables people to participate in online activities
  - Networking, tagging and commenting
- It shatters the barrier for online users to create and share information at any place at any time

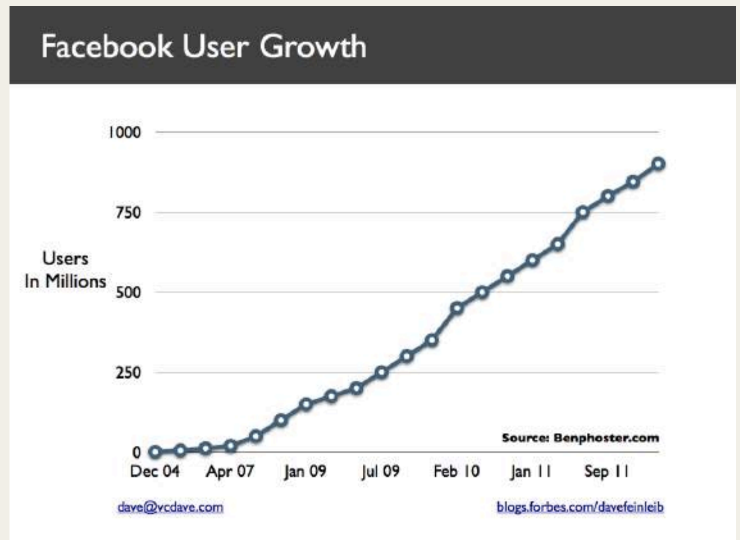


http://www.marketingprofs.com/charts/2010/4101/social-media-brand-followers-hunting-for-deals

Information Overload

- Social media data increases at an unprecedented rate

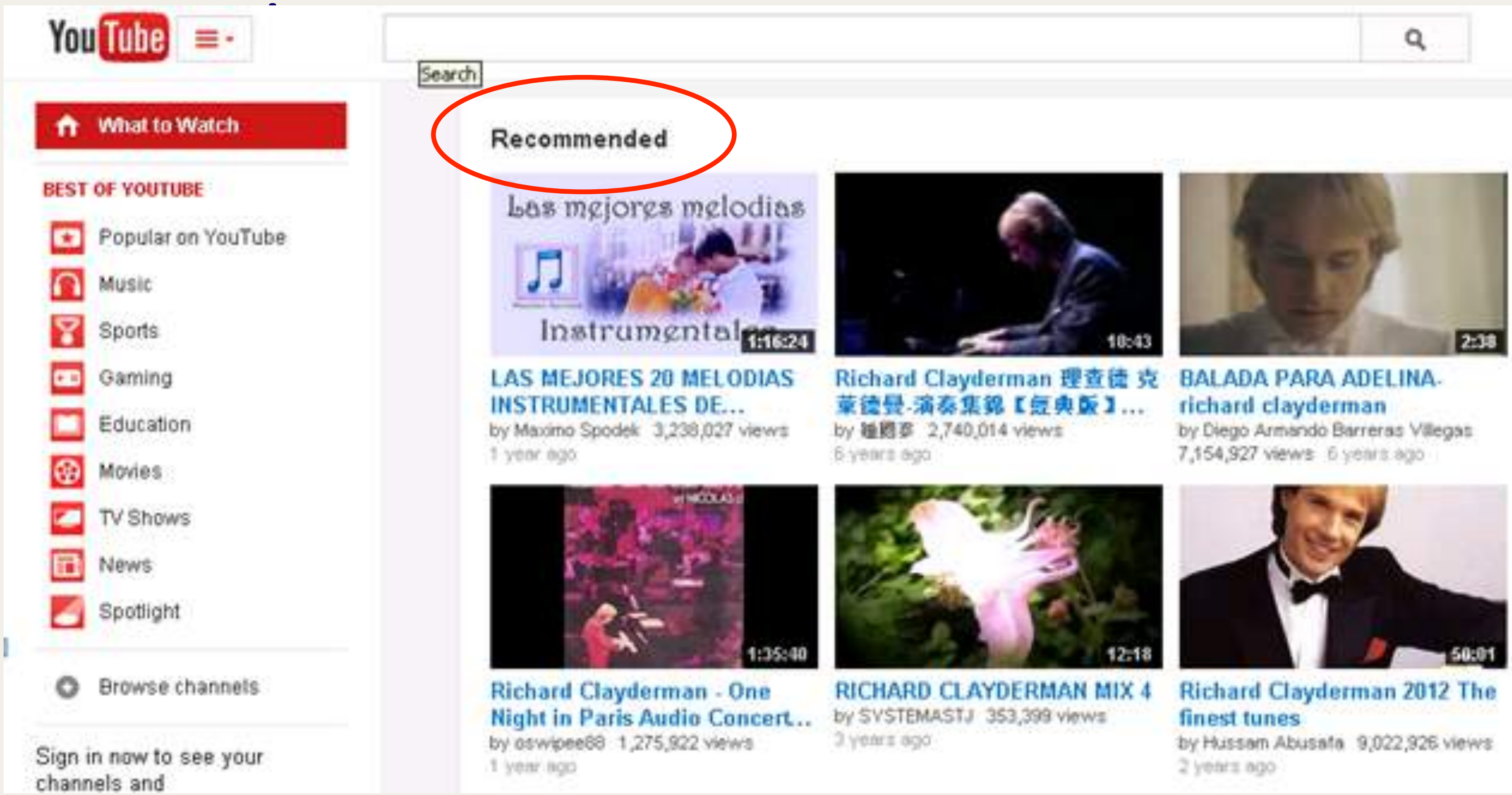




- It becomes increasingly difficult for online users to get their interested information

YouTube

- 100 hours of videos are uploaded into YouTube in

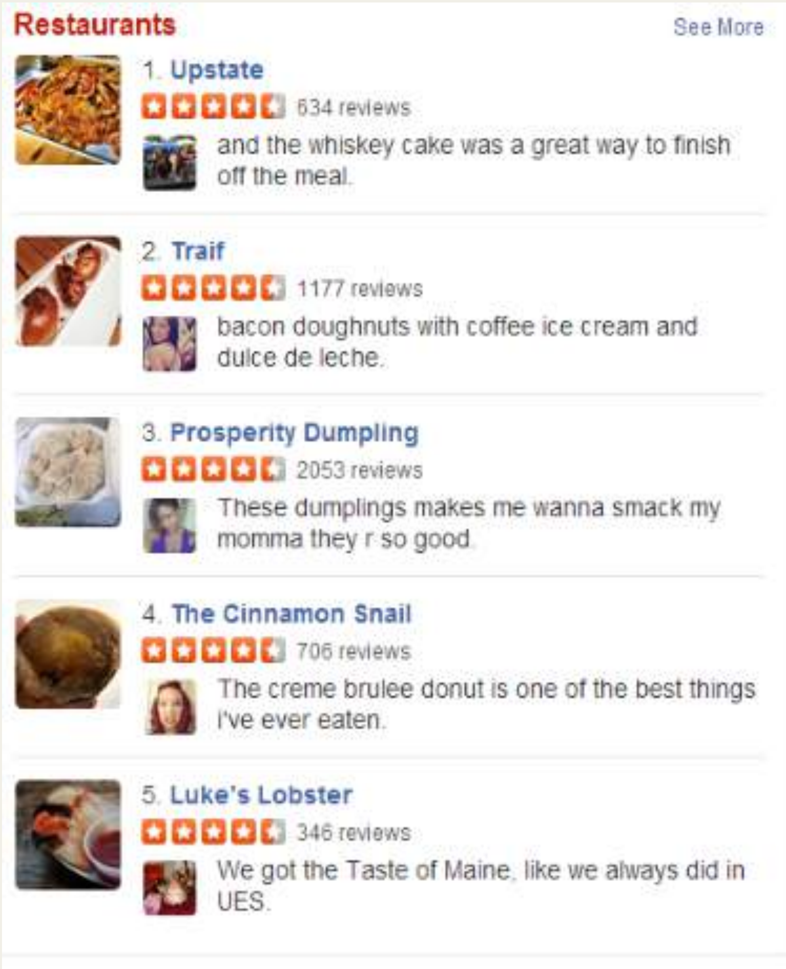


-

Yelp

- Among hundreds of thousands of restaurants in New York City, which one I should go for dinner?

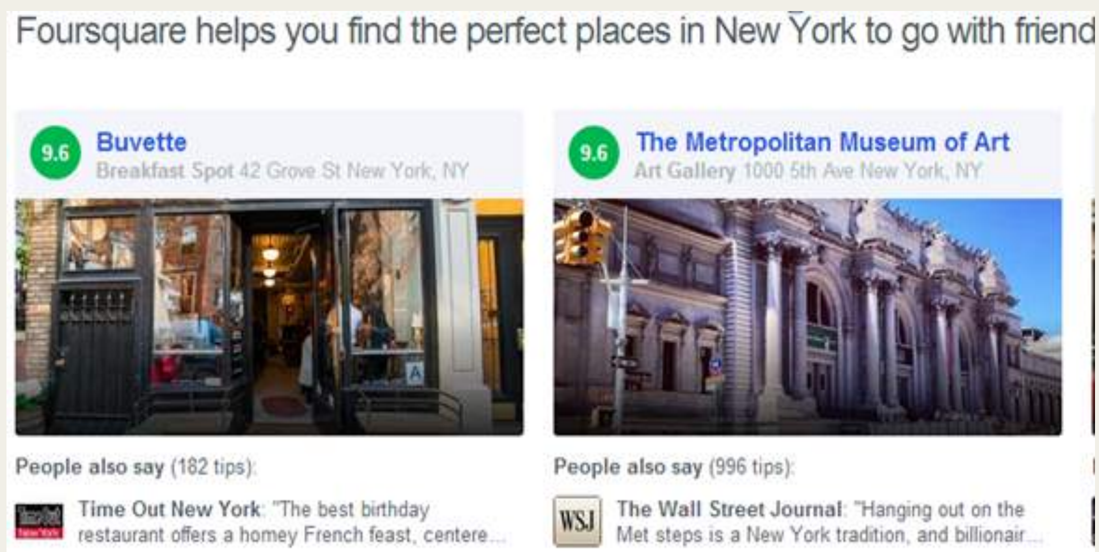
- Yelp can suggest some restaurants based on their ratings and your current locations automatically



Foursquare

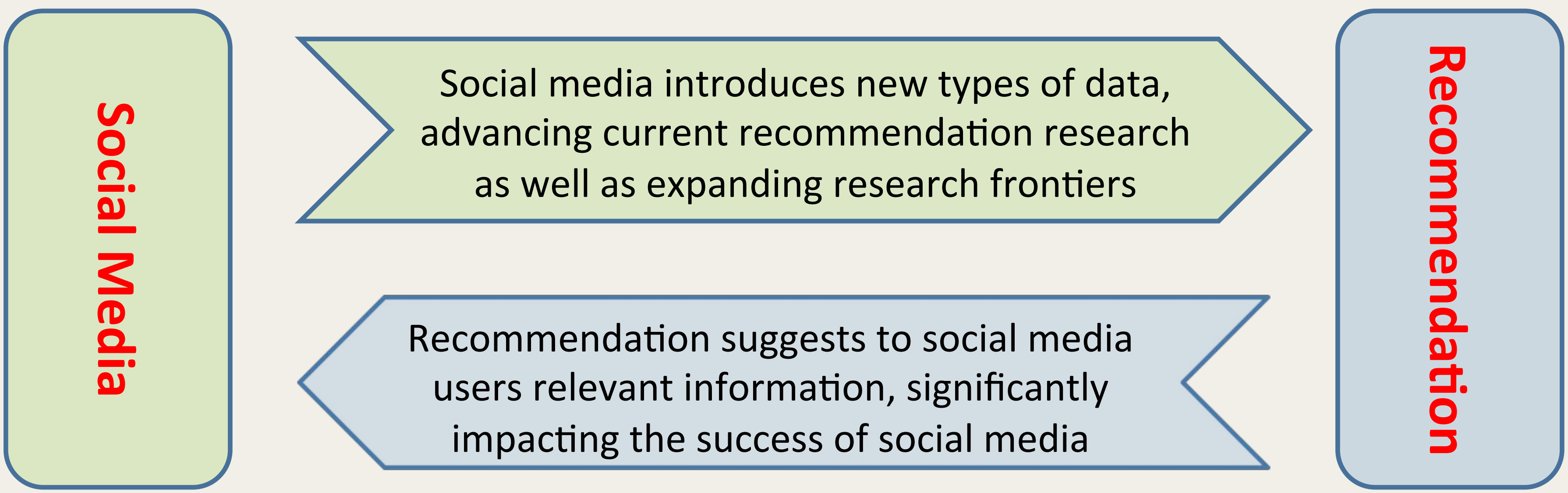
- During a short visit in New York City, where should we go?

- Foursquare can suggest some places to visit and as some useful tips base on your locations automatically



Recommendation and Social Media [Zafarani et al., 2014]

- Recommendation is widely used to mitigate *information overload* problem in social media
- Social media and recommendation can mutually benefit each other [**Guy and Carmel, 2011**]



Recommendation in Social Media [Tang et al., 2014]

- Social media users can be described with three types of information
  - Social information
  - Content information
  - Location information

- Three information types correspond to three recommendation tasks
  - Friend Recommendation
  - Content Recommendation
  - Location Recommendation

Special Characteristics of Recommendation

- Search starts with a user's explicit query
- Recommendation is triggered with a user's implicit query
- A user is provided with relevant and timely information without explicitly stating his needs
- Why is this necessary and critical?
  - If successful, a user will choose to stay longer on the site, or pay more attention, leading to more clicks, more purchases, more contributions, …

Recommendation in Social Media

Recommendation in Social Media

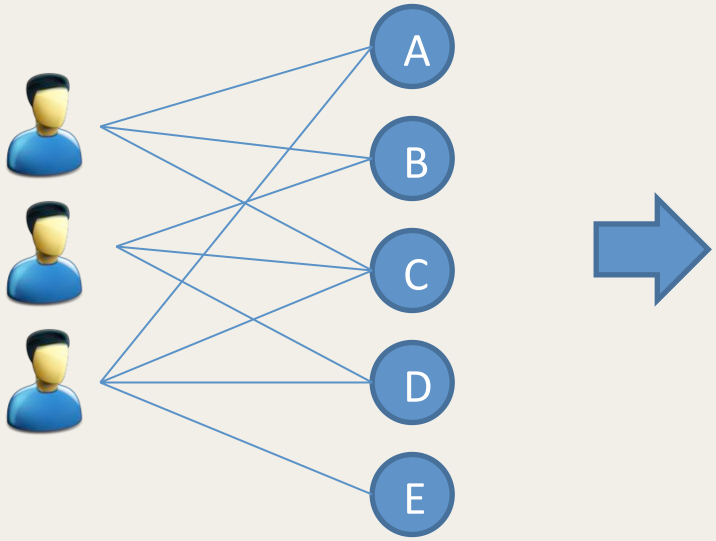
Content Recommendation in Social Media

Location Recommendation in Social Media

Performance Evaluations

Fundamental Recommendation Approaches [Adomavicius and Tuzhilin, 2005]

- User and content item relation can be represented as an user-item matrix **R**



	A	B	C	D	E
1	5	3	4	?	?
2	?	3	4	4	?
3	1	?	2	2	5

- Content-based recommendation
  - Recommending items similar to the ones that the user has preferred in the past
- Collaborative filtering (CF) – based recommendation
  - Using the user's past behavior to uncover user preferences
  - Memory-based CF and Model-based CF

Memory-based Collaborative Filtering

- It uses either the whole user-item matrix or a sample to generate a prediction
  – Needing memory to store the user-item matrix
- User-oriented collaborative filtering
  – Calculating user-user similarity
  – Aggregating ratings from similar users
- Item-oriented collaborative filtering
  – Computing item-item similarity
  – Aggregating ratings from similar items

User-oriented collaborative Filtering

- Calculating user-user similarity
  - –Cosine similarity

$$S(i,j) = \frac{\sum_{k\in I} R_{ik} R_{jk}}{\sqrt{\sum_{k\in I} R_{ik}^2}\sqrt{\sum_{k\in I} R_{jk}^2}}$$

- I is the set of items rated by u_i and u_j
- R_{ik} is the rating to the *k*th item from u_i

- Aggregating ratings from similar users

$$\hat{R}_{ij} = \frac{\sum_{u_k \in N_i} S_{ik} R_{kj}}{\sum_{u_k \in N_i} S_{ik}}$$

N_i is the set of users who have rated the j-th item

An Illustration of User-oriented Collaborative Filtering

	A	B	C	D	E
1	5	3	4	?	?
2	?	3	4	4	?
3	1	?	2	2	5

1, 2, and 3 are users
A, B, C, D, and E are items
R(1,D) = ?

- Calculating cosine similarity

$$S(1,2) = \frac{3*3+4*4}{\sqrt{3*3+4*4}\sqrt{3*3+4*4}} = 1$$

$$S(1,3) = \frac{5*1+4*2}{\sqrt{5*5+4*4}\sqrt{1*1+2*2}} = 0.9080$$

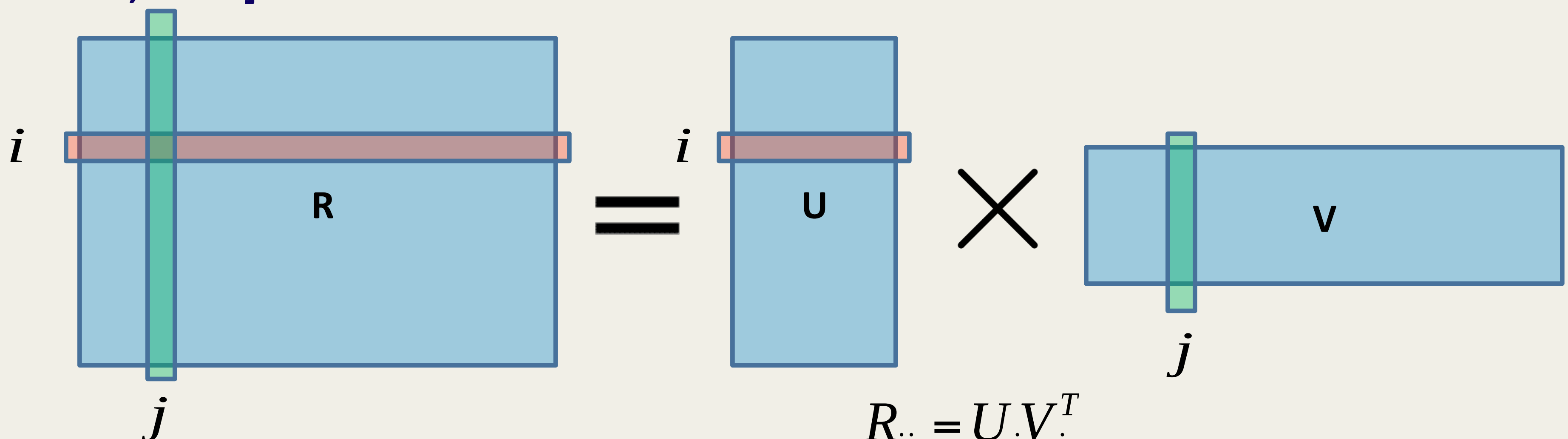
- Aggregating ratings

$$\hat{R}(1,D) = \frac{R(2,D)*S(1,2)+R(3,D)*S(1,3)}{S(1,2)+S(1,3)}$$

$$= \frac{4*1+2*0.9080}{1+0.9080} = 3.05$$

Model-based Collaborative Filtering

- It assumes there exists a model that generates the ratings and the model parameters can be learned
  - Storing only parameters instead of the rating matrix
  - Using the assumed model with parameters to do prediction
- Matrix factorization methods are very competitive and are widely adopted to build recommender systems **[Koren et al., 2009]**



$$R_{ij} = U_i V_j^T$$

An Illustration of Matrix Factorization based CF

	A	B	C	D	E
1	5	3	4	?	?
2	?	3	4	4	?
3	1	?	2	2	5

1,2 and 3 are users

A, B, C, D, and E are items

The latent dimension k =1

R(1,D) = ?

- Learning Latent Factors U and V

$$U = \begin{matrix} 2.6308 \\ 2.4109 \\ 1.4706 \end{matrix} \qquad V = \begin{matrix} 1.6252 \\ 1.2182 \\ 1.5740 \\ 1.5990 \\ 2.5716 \end{matrix}$$

- Reconstructing the rating matrix

$$\hat{R} = UV^T = \begin{matrix} 4.2756 & 3.2047 & 4.1408 & 4.2066 & 6.7654 \\ 3.9182 & 2.9368 & 3.7946 & 3.8550 & 6.1999 \\ 2.3901 & 1.7915 & 2.3147 & 2.3515 & 3.7819 \end{matrix}$$

Content Recommendation in Social Media

- New types of data introduced by social media have greatly enriched the sources available for content recommendation
  - Social information
  - Location information
- Content recommendation with social networks
  - How to include social information in content recommendation?
- Location-aware content recommendation
  - Given the locations of users, how to recommend their interested content?

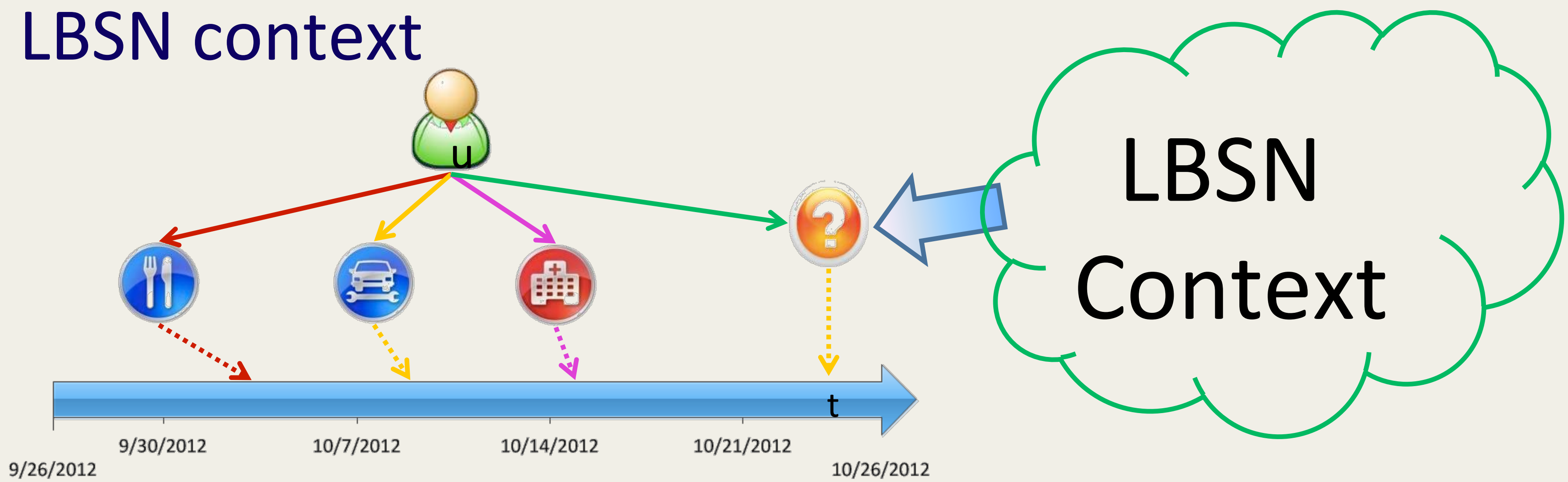
Recommendation in Social Media

Recommendation in Social Media

- Content Recommendation in Social Media
- Location Recommendation in Social Media
- **Performance Evaluations**

Location Recommendation in Social Media

- A number of location-based social networking services have emerged in recent years
  - Foursquare, Yelp, and Facebook Places
- Location recommendation is to recommend to a user some POIs for his future visits based on his LBSN context



Recommendation in Social Media

Recommendation in Social Media

- Content Recommendation in Social Media
- Location Recommendation in Social Media
- Performance Evaluations

Recommendation Evaluations

- Different evaluation metrics assess recommender systems from different perspectives
  - Prediction power: the ability to accurately predict users' choices
  - Classification accuracy: the ability to differentiate relevant items from irrelevant ones
  - Novelty and exploration: the ability to discover new items or explore diverse items

Prediction Accuracy Evaluation

- Prediction Accuracy Evaluation measures the average error of predicted ratings
  - Mean Absolution Error (MAE)

Predicted Ratings

Ground-truth Ratings

Testing Set

$$MAE = \frac{\sum_{<u_i, v_j> \in O} |\hat{R}_{ij} - R_{ij}|}{|O|}$$

  - Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{<u_i, v_j> \in O} (\hat{R}_{ij} - R_{ij})^2}{|O|}}$$

Ranking Accuracy Evaluation

- Ranking Accuracy evaluates how many recommended items are acquired by the users
- Precision@N
  - How many top-N recommended items are acquired?
  - For a target user u_i

$$Precision@N = \frac{|TopN(i) \cap L(i)|}{|TopN(i)|}$$

The items u_i acquired

The top-N items recommended to u_i

- Recall@N
  - How many top-N acquired items are recommended?
  - For a target user u_i

$$Recall@N = \frac{|TopN(i) \cap L(i)|}{|L(i)|}$$

Coverage Evaluation

- Item coverage
  - Evaluating how good the items recommended by a recommendation system S are

$$I_c = \frac{|N_d|}{|N|}$$

N is the set of items supposed to be recommended, while Nd is the set of items recommended by S

- User coverage
  - Evaluating how good the users recommended by a recommendation system S are

$$U_c = \frac{|M_d|}{|M|}$$

M is the set of users supposed to be recommended, while Md is the set of users S recommends

References

[**Zafarani et al., 2014**]R. Zafarani, M. Abbasi, and H. Liu. Social Media Mining: An Introduction. Cambridge University Press, 2014.

[**Guy and Carmel, 2011**] I. Guy and D. Carmel. Social recommender systems. In Proceedings of the 20th international conference companion on World wide web, pages 283–284. ACM, 2011.

[**Tang et al., 2014**] J. Tang, J. Tang, and H. Liu. Recommendation in Social Media: Recent Advances and New Frontiers. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp, 1977-1977, 2014.

[**Liben-Nowell and Kleinberg, 2007**] D. Liben-Nowell and J. Kleinberg. link-prediction problem for social networks. Journal of the American society for information science and technology, pp, 1019—1031, 2007.

[**Adomavicius and Tuzhilin, 2005**] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. Knowledge and Data Engineering, IEEE Transactions on, 17(6):734–749, 2005

[**Koren et al., 2009**] Y. Koren, R. Bell, and C. Volinsky. factorization techniques for recommender systems. Computer, pp, 30—37, 2009.

Outline

Introduction

Content Recommendation

Location Recommendation

Future Work

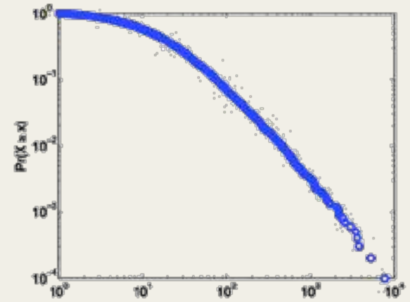
The Scope of Content Recommendation in the Tutorial

- Content may be manifested in diverse ways such as tweets, images, music, products, or videos
  - We do not assume that an item-feature matrix is available
  - User-item relations can be represented as a user-item matrix
- We only focus on collaborative filtering algorithms
  - Widely used
  - Promising performance in many real-world recommender systems

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	5	2		3		4		
u_2	4	3			5			
u_3	4		2				2	4
u_4								
u_5	5	1	2		4	3		
u_6	4	3		2	4		3	5

Challenges of Traditional Approaches

- Data sparsity problem
  - –Content in social media is big but the available content for most individuals is very limited
  - –The user-item matrix is extremely sparse with less than 1% observed entities



- Cold-start users
  - –The number of entities per user follows a power-law distribution
  - –Many users have no or few entities

Opportunities from Social Media

- Social media provides additional sources for content recommendation
  – Social information and location information
  – Mitigating data sparsity problem

- We may make recommendations for cold-start users based on other information sources
  – Users' preferences are similar to their networks
  – Reducing significantly the number of cold-start users

Content Recommendation

Content Recommendation

Content Recommendation with Social Networks

Location-aware Content Recommendation

Content Recommendation

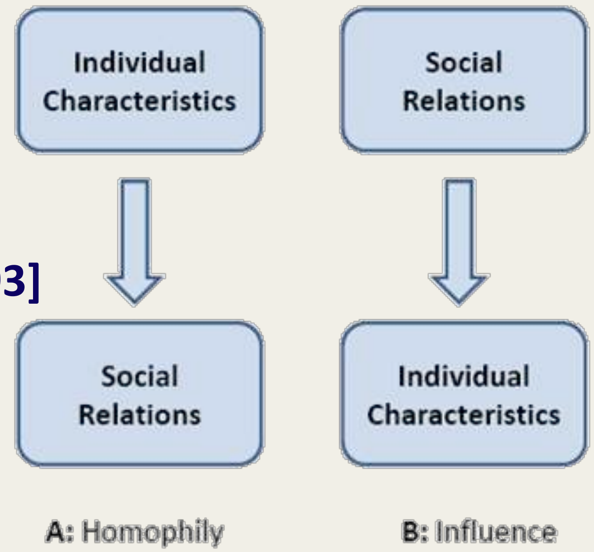
Content Recommendation

Content Recommendation with Social Networks

Location-aware Content Recommendation

Why Social Networks

- Social networks provide complementary information
  - Overlap between one's similar users and her social network is less than 10% **[Crandall et al., 2009]**
- Users' preferences are likely to similar to their networks
  - Homophily **[McPherson et al., 2001]**
  - Social influence **[Marsden and Friedkin, 1993]**



Categorization of Social Recommendation [Tang et al., 2013]

- Most existing social recommender systems are CF-based methods

CF-based Social Recommendation model	=	A basic CF model	+	A Social Model

- We can categorize social recommender systems based on their basic CF models
  - Memory-based social recommender systems
  - Model-based social recommender systems

Memory-based Social Recommendation

- It uses memory-based CF methods, especially user-oriented methods, as basic models
- It usually consists of two steps
  - Step 1: obtaining relevant users N_i for user i,
  - Step 2: aggregating recommendations from N_i
- Different systems in this category provide different ways to obtain relevant users in Step 1

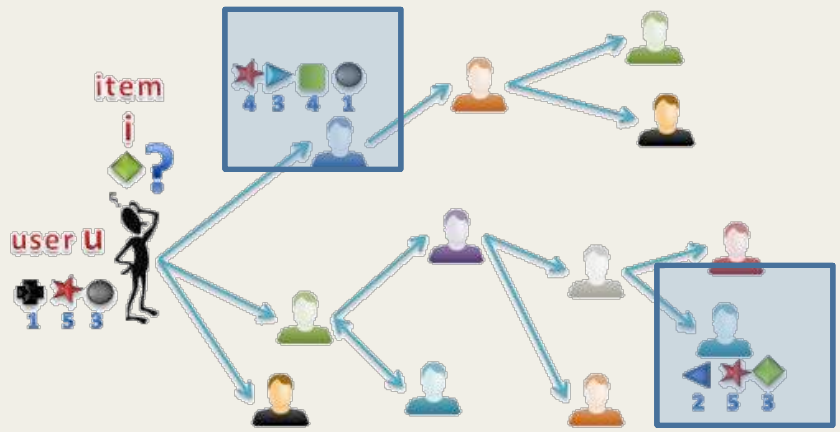
TidalTrust and MoleTrust

- TidalTrust only considers users at the shortest distance [**Golbeck, 2005**]
  - Efficient
  - High precision
  - Low recall
- MoleTrust considers raters up to a maximum-depth *d* **[Massa and Avesani, 2004]**
  - Trade-off between precision and recall

$$\hat{\mathbf{R}}_{ij} = \bar{\mathbf{R}}_i + \frac{\sum_{u_k \in \mathcal{N}_i} \mathbf{S}_{ik}(\mathbf{R}_{kj} - \bar{\mathbf{R}}_k)}{\sum_{u_k \in \mathcal{N}_i} \mathbf{S}_{ik}}$$

TrustWalker [Jamali and Ester, 2009]

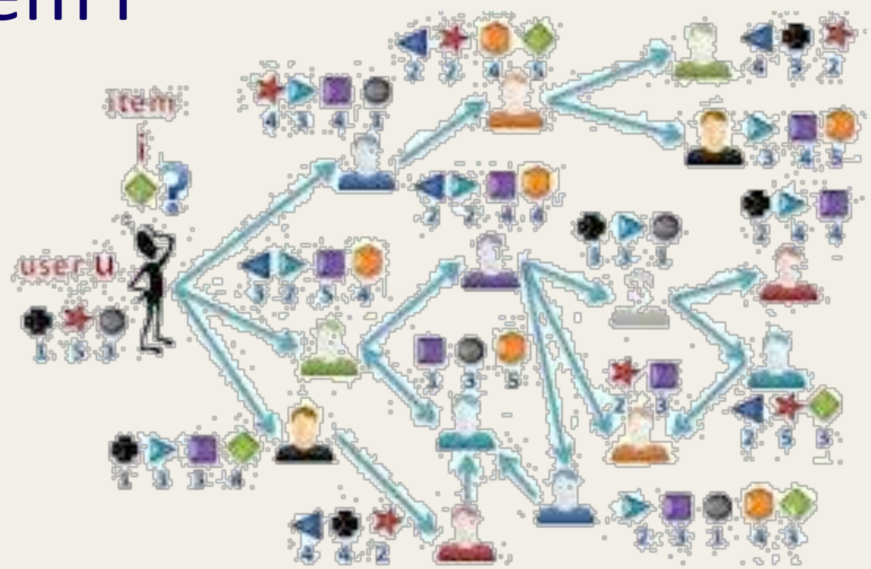
- In addition to distant users who have rated the target item, it also uses near friends who have rated similar items
  - Distant users on the exact target item
  - Close friends on similar items



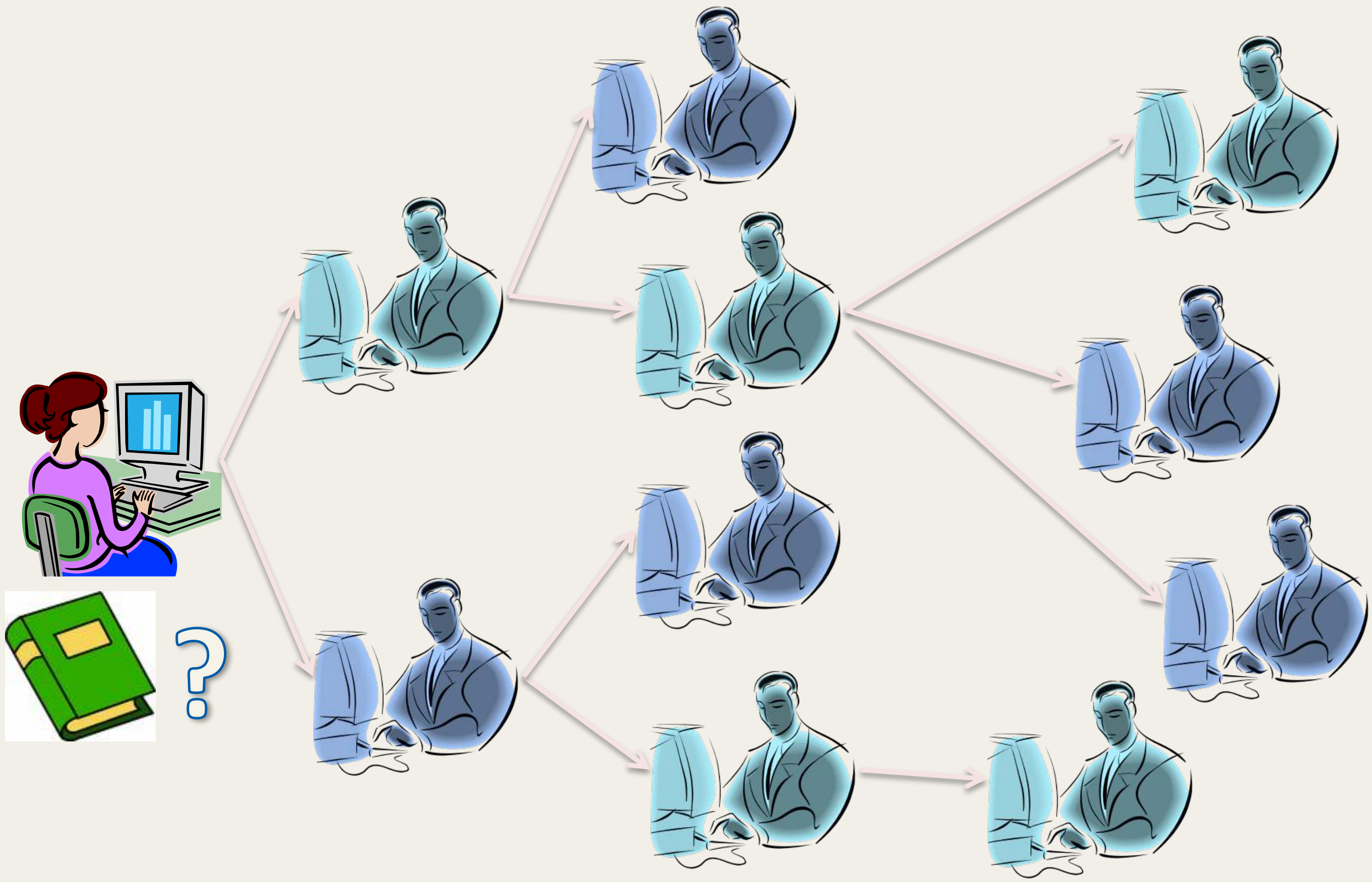
- It combines item-based recommendation and trust-based recommendation via random walk

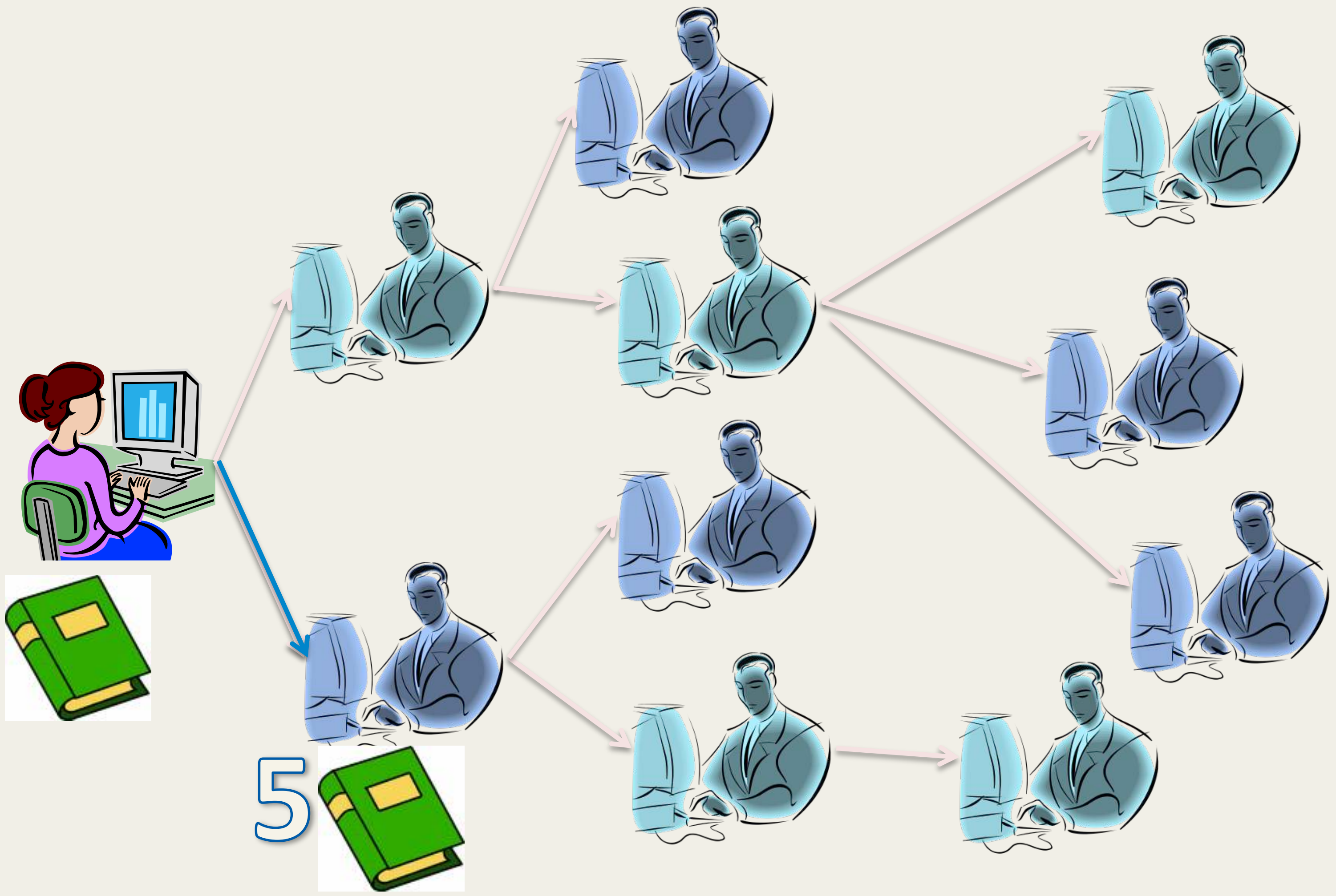
TrustWalker

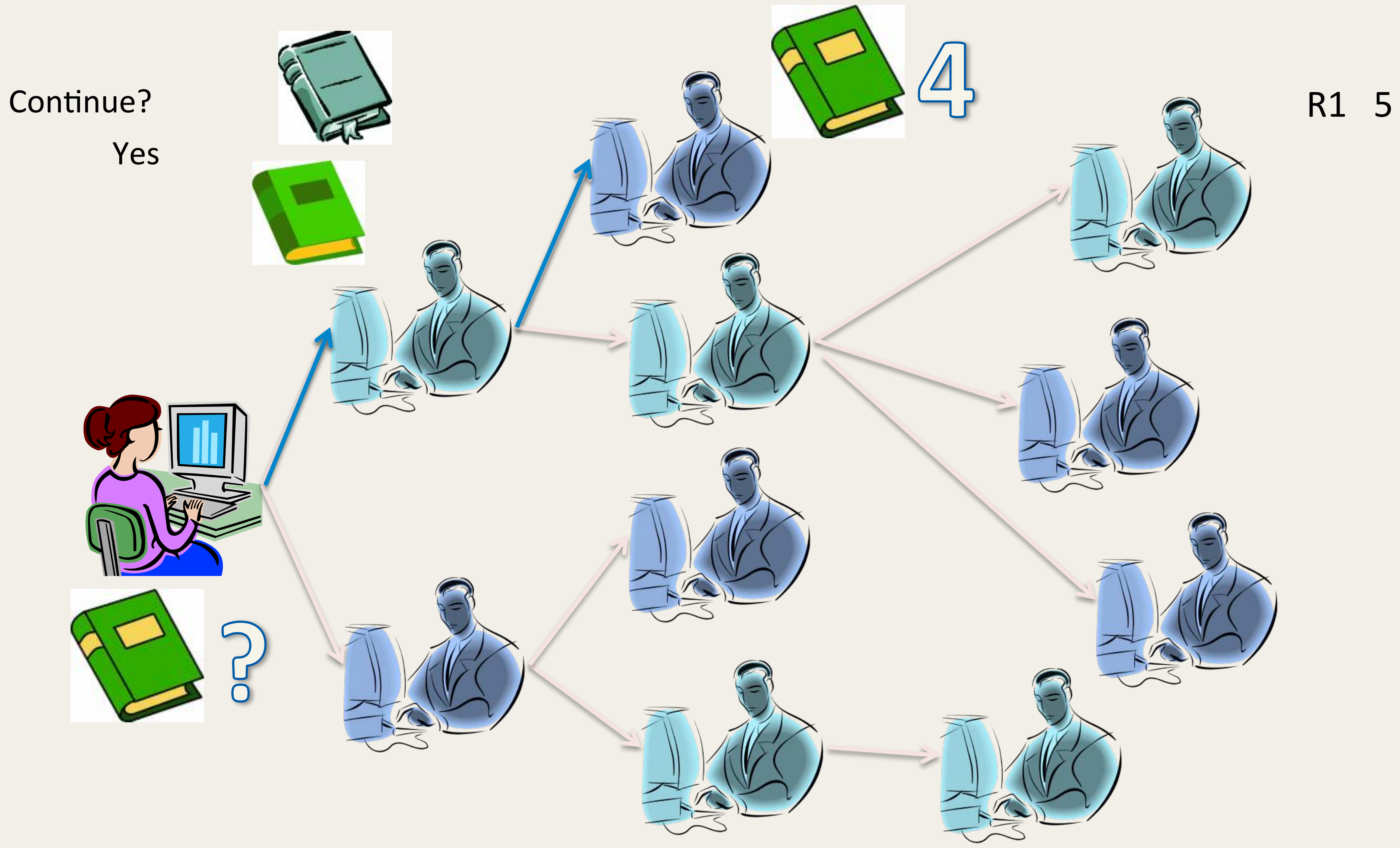
- Each random walk starts from a target user u to seek rating score for item i

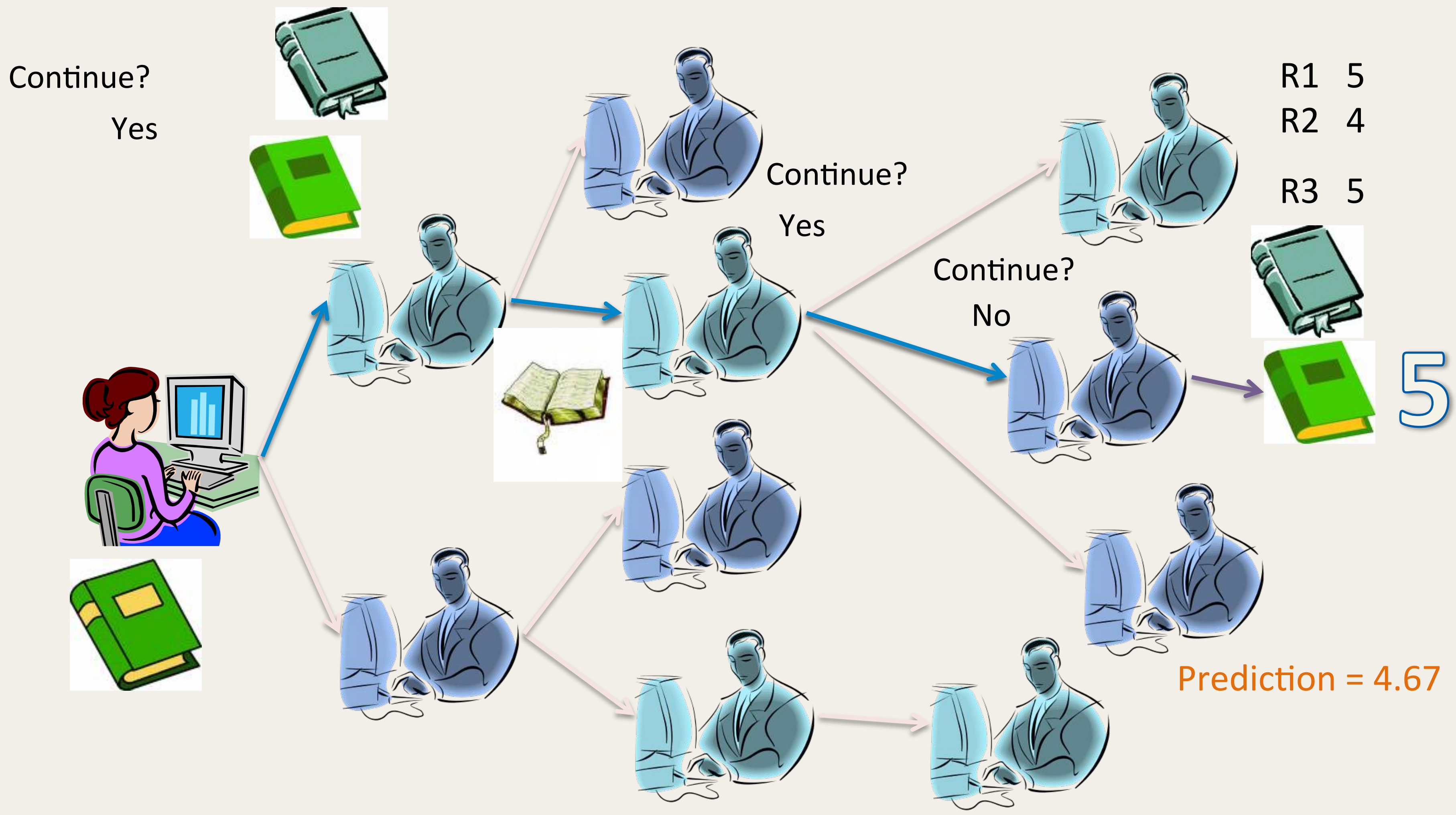


- In step *k* at user *v*:
  - If v has rated i, return R_{vi}
  - With the probability Q_{vik} , stop random walk, select a similar item j rated by u and return R_{vj}
  - With the probability 1 - Q_{vik}, continue the random walk to a direct neighbor of v




5


Continue?
Yes
4
R1 5
?


Continue?
Yes
Continue?
Yes
Continue?
No
R1 5
R2 4
R3 5
5
Prediction = 4.67

Model-based Social Recommendation

- In such systems, model-based CF methods are used
  - Matrix factorization is widely chosen as the basic model
- There are three common ways to integrate social information under the matrix factorization framework **[Tang et al., 2013]**
  - Co-factorization methods
  - Ensemble methods
  - Regularization methods

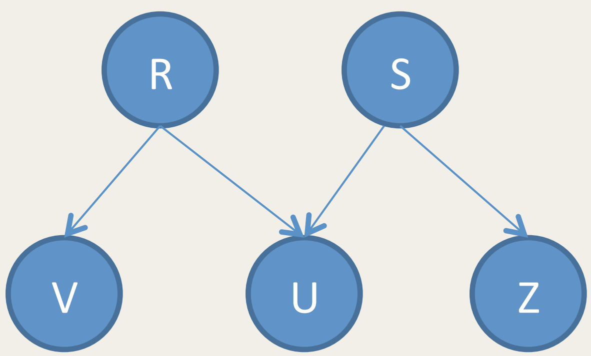
Co-factorization Methods

- Co-factorization methods perform co-factorization on the user-item matrix R and the user-user social matrix S

- SoRec **[Ma et al., 2008]**

$$R_{ij} = U_i V_j^T$$

$$S_{ij} = U_i Z_j^T$$



Ensemble Methods

- Ensemble methods combine recommendations for a user and her social network

Recommendation for u + *α* **Recommendation for u's social network**

- STE [Ma et al., 2009]

$$R_{ij} = \boxed{U_i V_j^T} + \alpha \boxed{\sum_{u_k \in N(u_i)} S_{ik} U_k V_j^T}$$

Regularization Methods

- Regularization methods add a regularization term to force users' preferences to be close to those of their social networks

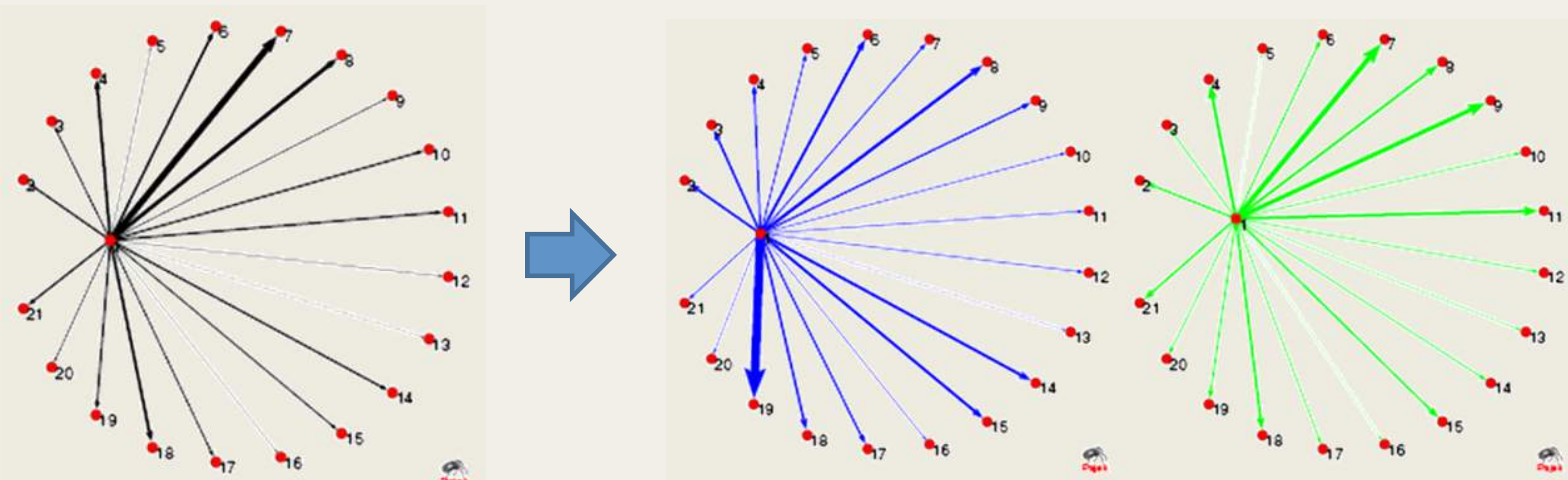
- SocialMF [Jamali and Ester, 2010]

$$\min \sum_i \| U_i - \boxed{\sum_{u_k \in N(u_i)} S_{ik} U_k} \|$$

The average user preference of the social network of u_i

Handling Heterogeneity

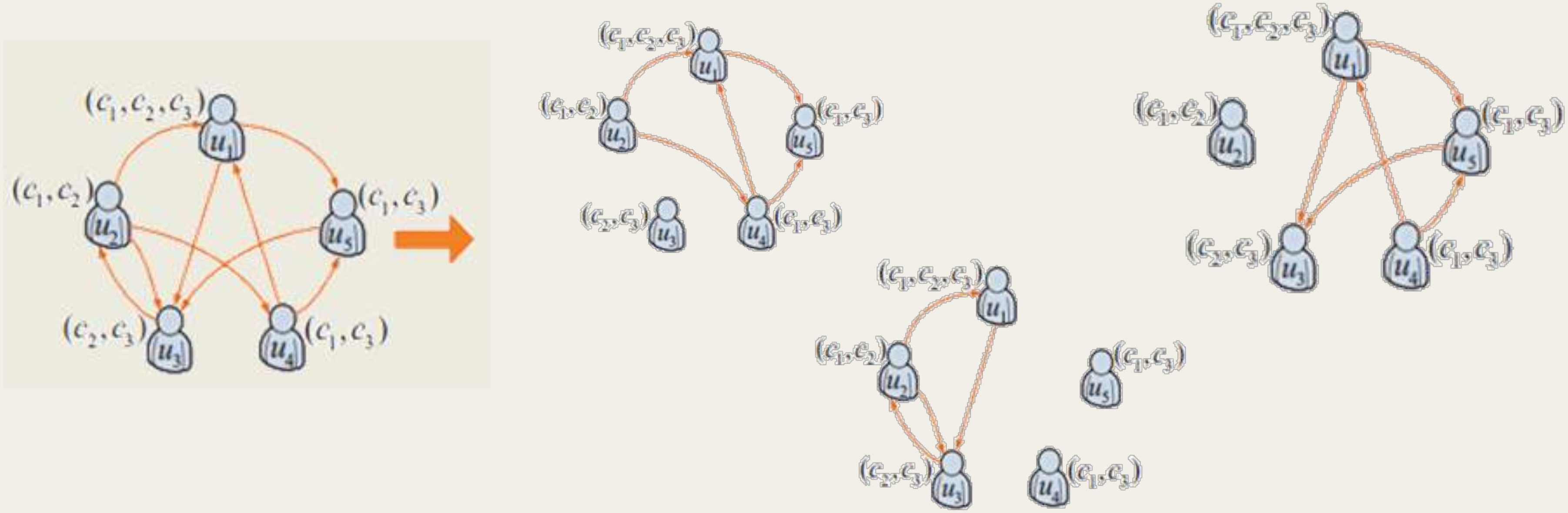
- Users trust their friends differently in different domains [Tang et al. 2012]



(a) Single Trust (b) Trust in Home & Garden (c) Trust in Restaurants

Circle-based Social Recommendation [Yang et al. 2012]

- Trust circle inference
  - v is in inferred circle c of u iff u connects to v and both of them are interested in the category c

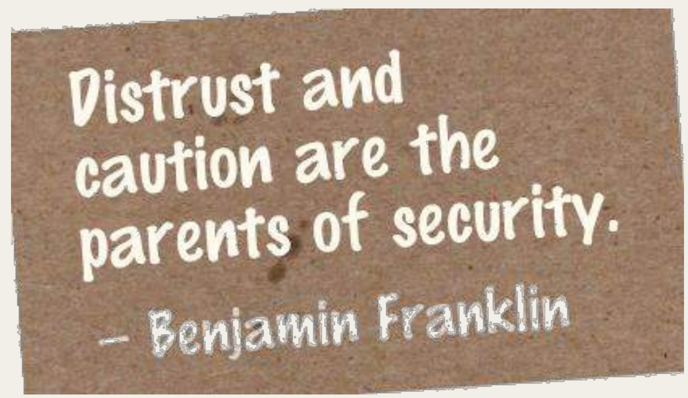


- SocialMF is applied to make recommendations for each circle

Distrust in Social Media [Tang, 2015]

- Distrust tends to be more noticeable and credible, and weighed more in decision making than trust

- Distrust is not the negation of trust and has significant added value
  - A small amount of distrust information can make remarkable improvement in link prediction



Distrust in Social Recommendation [Victor et al., 2009]

- Distrust as a filter
  - Using distrust to filter out ``unwanted'' users in the recommendation processes

$$\hat{r}_{u,i} = \bar{r}_u + \frac{\boxed{\sum_{v \backslash \mathcal{D}}} (r_{v,i} - \bar{r}_v) \times t_{u,v}}{\sum t_{u,v}}$$

- Distrust as a dissimilarity measure

$$\hat{r}_{u,i} = \bar{r}_u + \frac{\sum_{v \backslash \mathcal{D}} (r_{v,i} - \bar{r}_v) \times t_{u,v}}{\sum t_{u,v}} \boxed{- \frac{\sum_{v \in \mathcal{D}} (r_{v,i} - \bar{r}_v) \times d_{u,v}}{\sum d_{u,v}}}$$

Is Distrust Dissimilarity? [Tang, 2015]

- Distrust is not a dissimilarity measurement
  - CI: Commonly-rated Items
  - COSINE: Rating-cosine similarity
  - COSINE-CI: Rating-cosine similarity of commonly rated items

	CI	COSINE	COSINE-CI
Distrust ($\mathbf{s}_d$)	0.4994	0.0105	0.0142
Trust $\mathbf{s}_t$	0.6792	0.0157	0.0166
Random Pairs ($\mathbf{s}_r$)	0.1247	0.0027	0.0032

- Using distrust for recommendation is still an open challenge

Content Recommendation

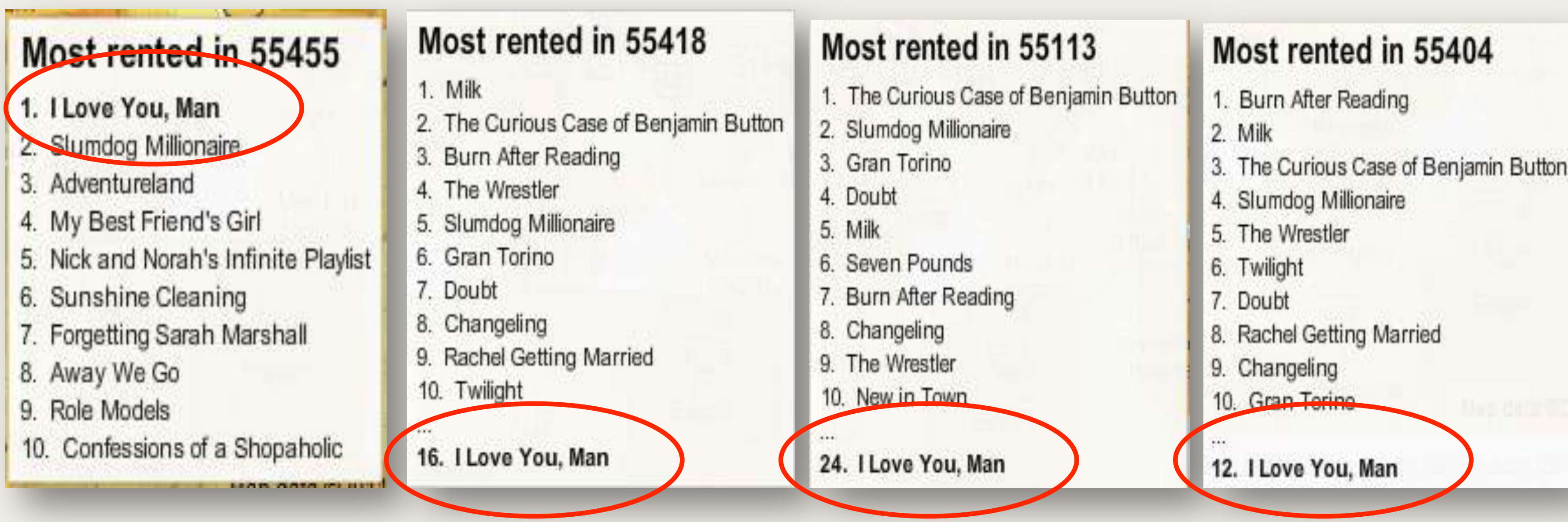
Content Recommendation

Content Recommendation with Social Networks

Location-aware Content Recommendation

Why Users' Locations Matter ?

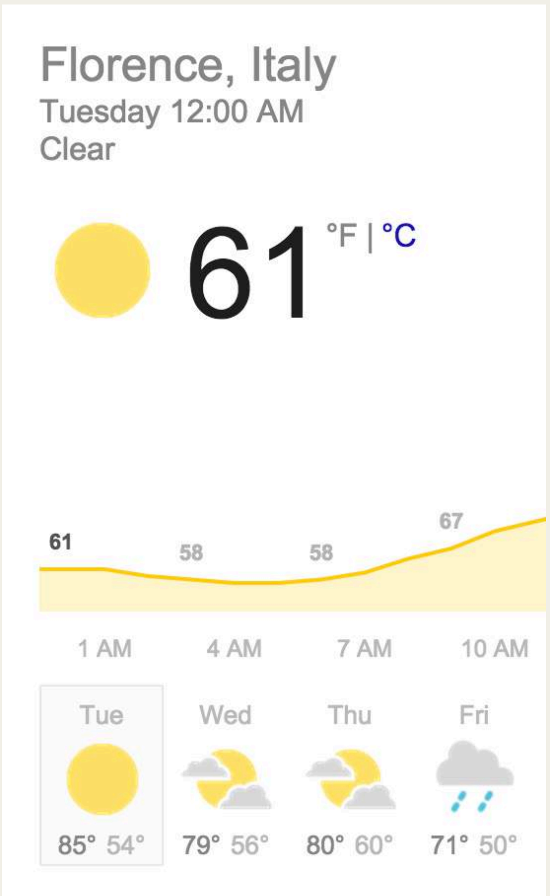
- Users' preferences may differ based on the user locations
  - New York Times has a very interesting visualization tool for Netflix rental patterns by zip code

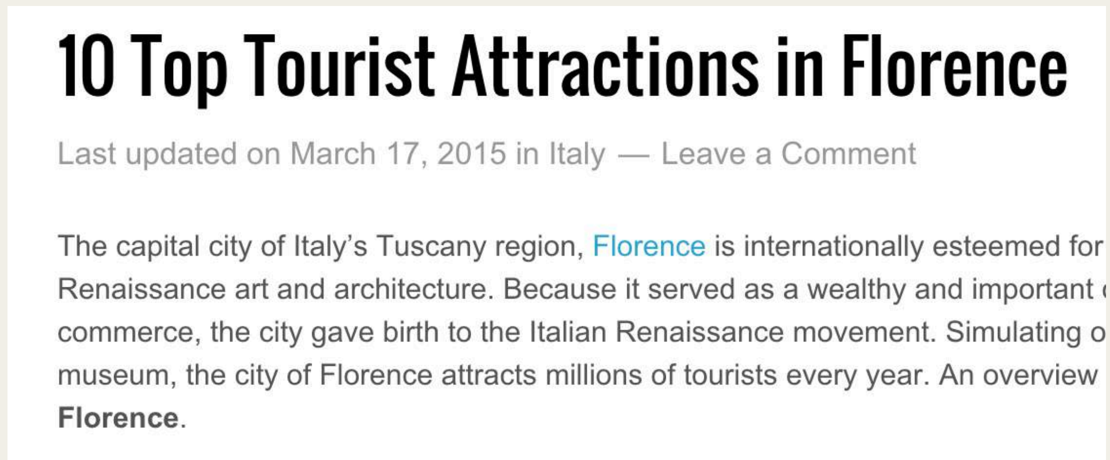


http://www.nytimes.com/interactive/2010/01/10/nyregion/20100110-netflix-map.html?_r=0

Why Users' Locations Matter?

- Users are more interested in content that is close to their current locations





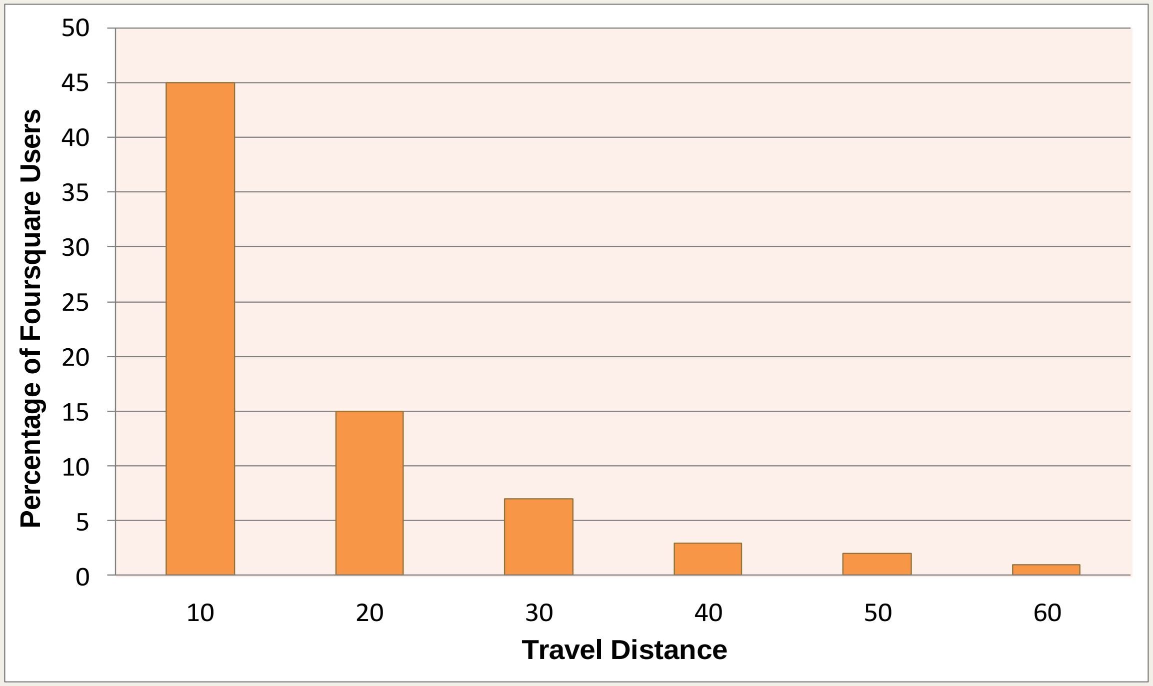
10 Top Tourist Attractions in Florence

Last updated on March 17, 2015 in Italy — Leave a Comment

The capital city of Italy's Tuscany region, Florence is internationally esteemed for Renaissance art and architecture. Because it served as a wealthy and important commerce, the city gave birth to the Italian Renaissance movement. Simulating o museum, the city of Florence attracts millions of tourists every year. An overview **Florence**.

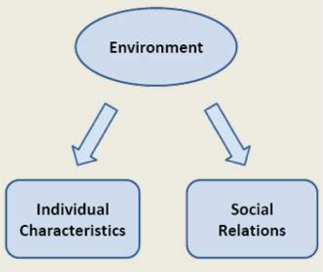
Why Users' Locations Matter?

- In terms of locations, users tend to travel a limited distance
  - 75 % of users travel less than 50 miles



Location-aware Content Recommender Systems

- Location Distance Weighted Methods
  – Confounding effects

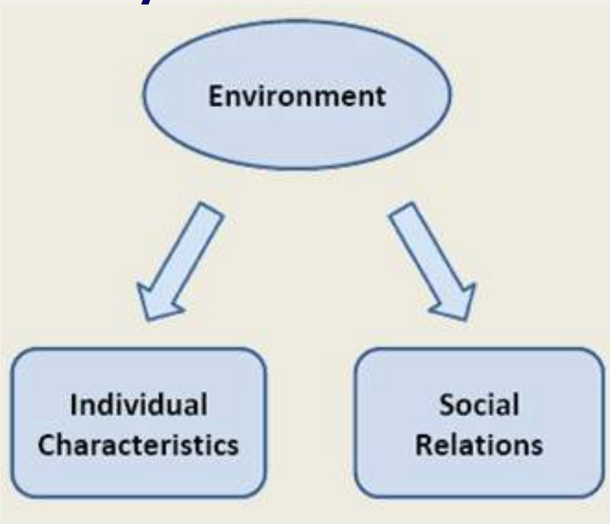


- User-partition Based Methods
  – Partition users based on their locations

- Item-partition Based Methods
  – Partition items based on their associated locations

Location Distance Weighted Methods [Yue et al. 2013]

- Geographically closed users are likely to share similar user preferences
  - Confounding



- Calculating location similarity as

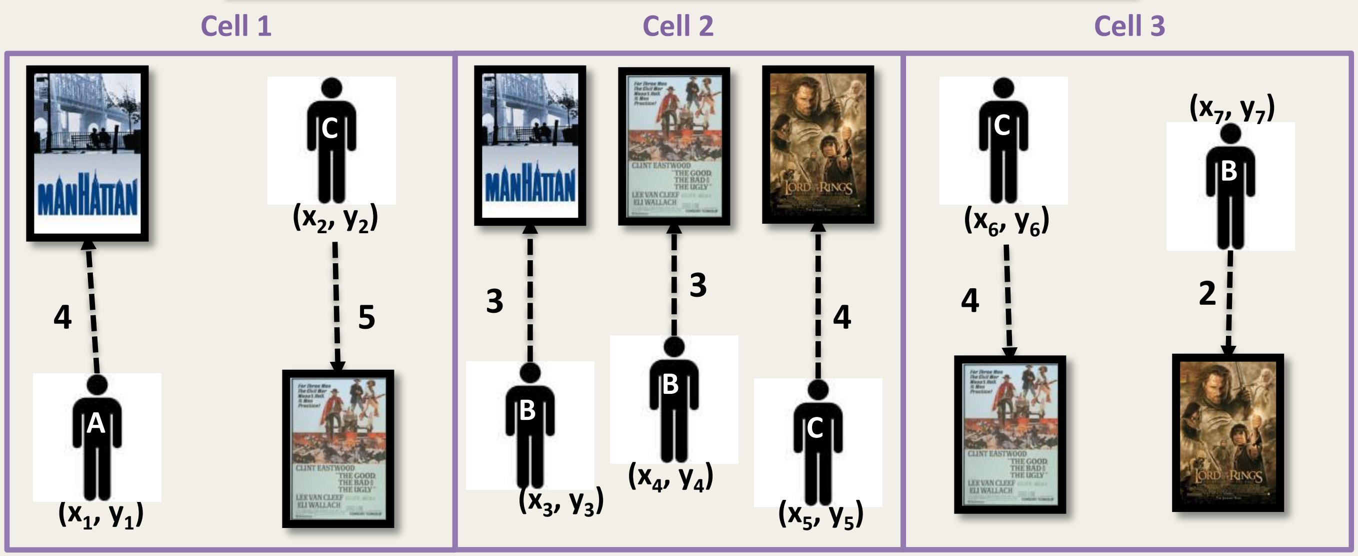
$$L_{uv} = \frac{1}{1 + \alpha * distance(u, v)}$$

- Incorporating location similarity into user-oriented collaborative filtering

$$R_{ui} = \sum_{v \in N_u} L_{uv} w_{uv} R_{vi}$$

User Partition Based Methods [Levandoski et al., 2012]

1. Partition ratings by user locations



2. Build collaborative filtering model for each cell using only ratings contained within the cell

Cell 1

Build Collaborative Filtering Model using:

User	Item	Rating
A		4
C		5

Cell 2

Build Collaborative Filtering Model using:

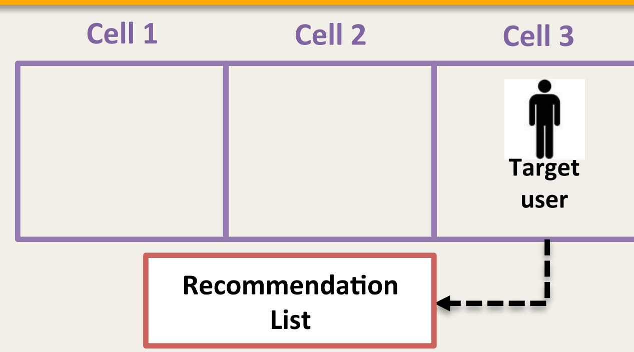
User	Item	Rating
B		3
B		3
C		4

Cell 3

Build Collaborative Filtering Model using:

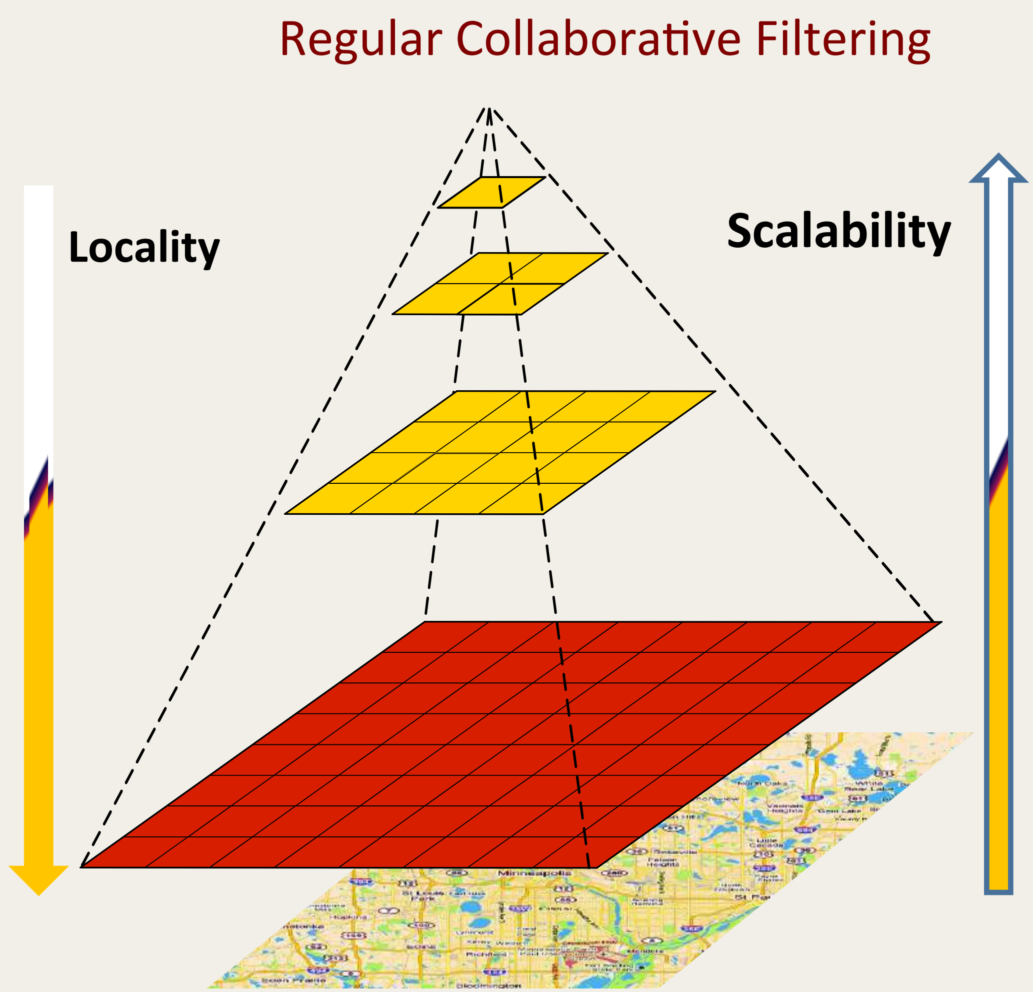
User	Item	Rating
B		4
C		5

3. Generate recommendations using collaborative filtering using the model of the cell containing the target user



User Partition Structure

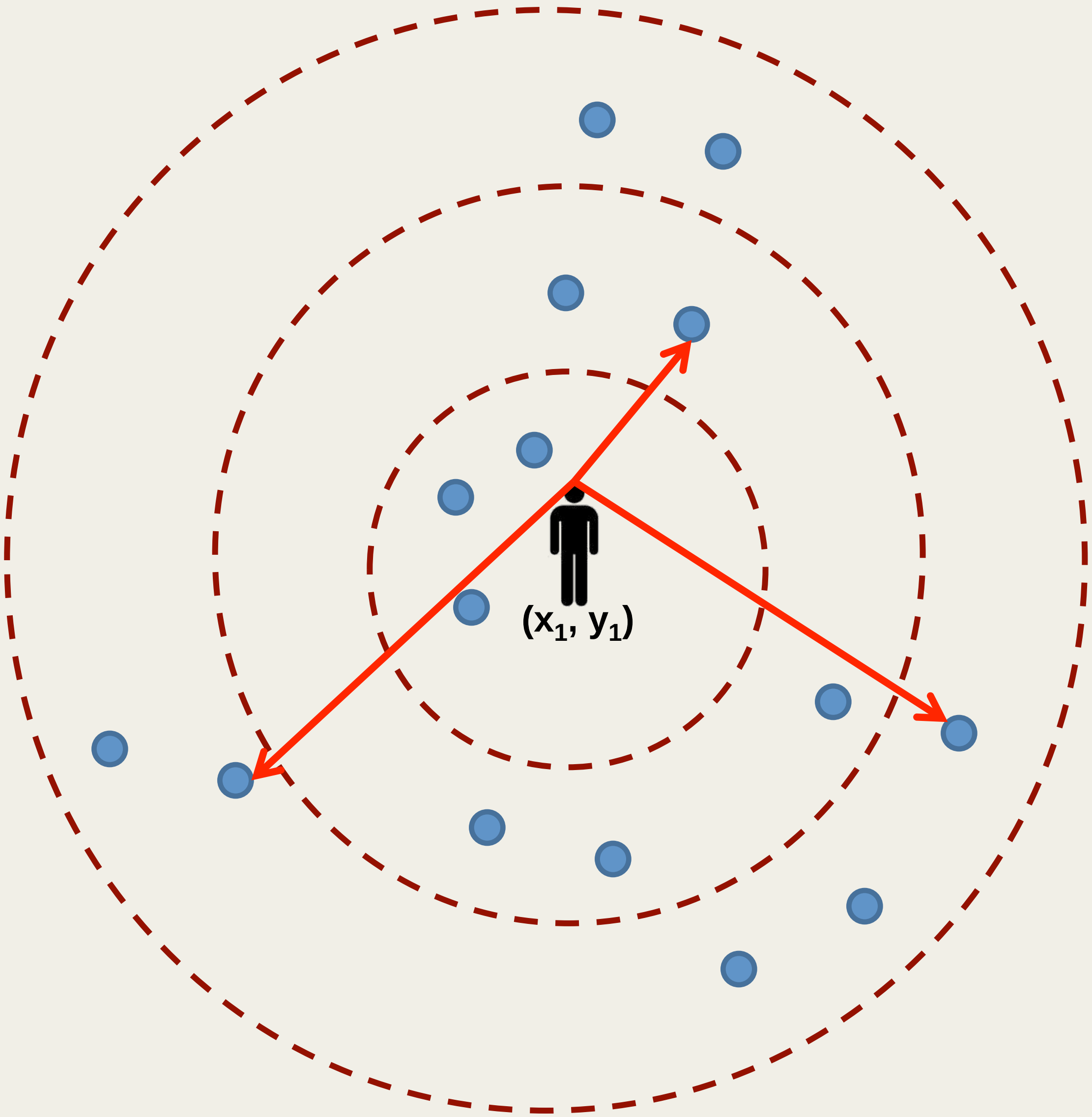
- Adaptive Pyramid Structure
  - Hotel Caravaggio, Florence, Tuscany, Italy, EU
- Two main goals:
  - Locality
  - Scalability



Smaller cells → more "localized" answers

Item-partition based Methods

- Partition items based on their associated locations
- Penalizing the item based on its distance from the user
- Recommending items within a certain distance



References

[McPherson et al., 2001] McPherson, M., Smith-Lovin, L., Cook, J. Birds of a feather: Homophily in social networks. Annual review of sociology pp. 415–444, 2001.

[Marsden and Friedkin, 1993] Marsden, P., Friedkin, N.: Network studies of social influence. Sociological Methods and Research 22(1), 127–151, 1993.

[Crandall et al., 2009] Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S. Feedback effects between similarity and social influence in online communities. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 160–168, 2008

[**Tang et al., 2013**] J. Tang, X. Hu, and H. Liu. Social recommendation: a review. Social Network Analysis and Mining, 3(4):1113–1133, 2013.

[**Golbeck, 2005**] J. A. Golbeck. Computing and applying trust in web-based social networks. 2005.

[Massa and Avesani, 2004] Massa, P., Avesani, P. Trust-aware collaborative filtering for recommender systems. In: On the Move to Meaningful Internet Systems, pp. 492–508, 2004.

[**Jamali and Ester, 2009**] M. Jamali and M. Ester. Trustwalker: a random walk model for combining trust-based and item-based recommendation. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 397–406. ACM, 2009.

References

[**Ma et al., 2008**] H. Ma, H. Yang, M. R. Lyu, and I. King. Sorec: social recommendation using probabilistic matrix factorization. In Proceedings of the 17th ACM conference on Information and knowledge management, pages 931–940, 2008.

[**Ma et al., 2009**] H. Ma, I. King, and M. R. Lyu. Learning to recommend with social trust ensemble. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pages 203–210, 2009.

[**Jamali and Ester, 2010**] M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In Proceedings of the fourth ACM conference on Recommender Systems, pages 135–142, 2010.

[**Tang et al. 2012**] J. Tang, H. Gao, and H. Liu. mTrust: discerning multi-faceted trust in a connected world. In Proceedings of the fifth ACM international conference on Web search and data mining, pages 93–102, *2012.*

[**Yang et al. 2012**] X. Yang and H. Steck, and Y. Liu. Circle-based recommendation in online social networks. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp, 1267—1275, 2012.

[**Tang, 2015**] J. Tang. Computing Distrust in Social Media. Ph.D. Dissertation, Arizona State University, 2015.

References

[**Victor et al., 2009**] P. Victor, C. Cornelis, M. De Cock, and P. Pinheiro da Silva. Gradual trust and distrust in recommender systems. Fuzzy Sets and Systems, 160(10), 1367–1382, 2009.

[**Yue et al. 2013**] W. Yue, M. Song, J.Han and H. E. Location context aware collective filtering algorithm. Pervasive Computing and the Networked World, pp, 788—800, 2013.

[**Levandoski et al., 2012**] J. Levandoski, S. Mohamed, and E. Ahmed, and M, Mohamed. Lars: A location-aware recommender system. IEEE 28th International Conference on Data Engineering, pp, 450—461, 2012.

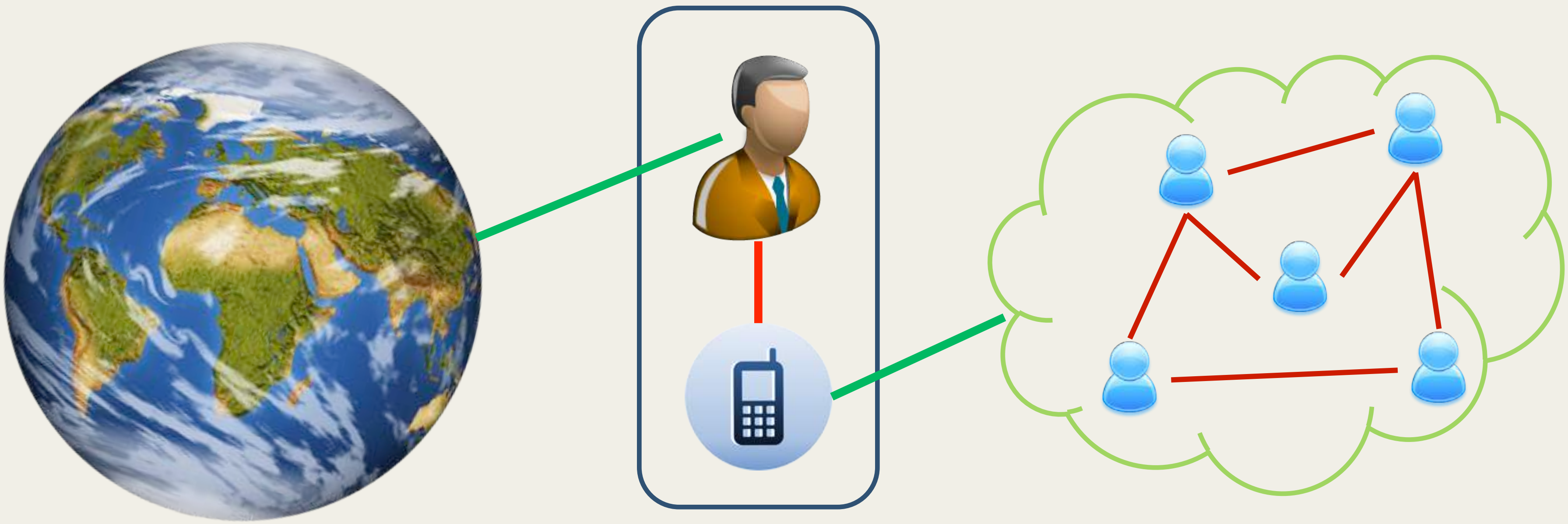
[**Bao et al., 2012**] J. Bao and M. Mokbel, and C. Chow. GeoFeed: A location aware news feed system. 2012 IEEE 28th International Conference on Data Engineering , 2012.

Outline

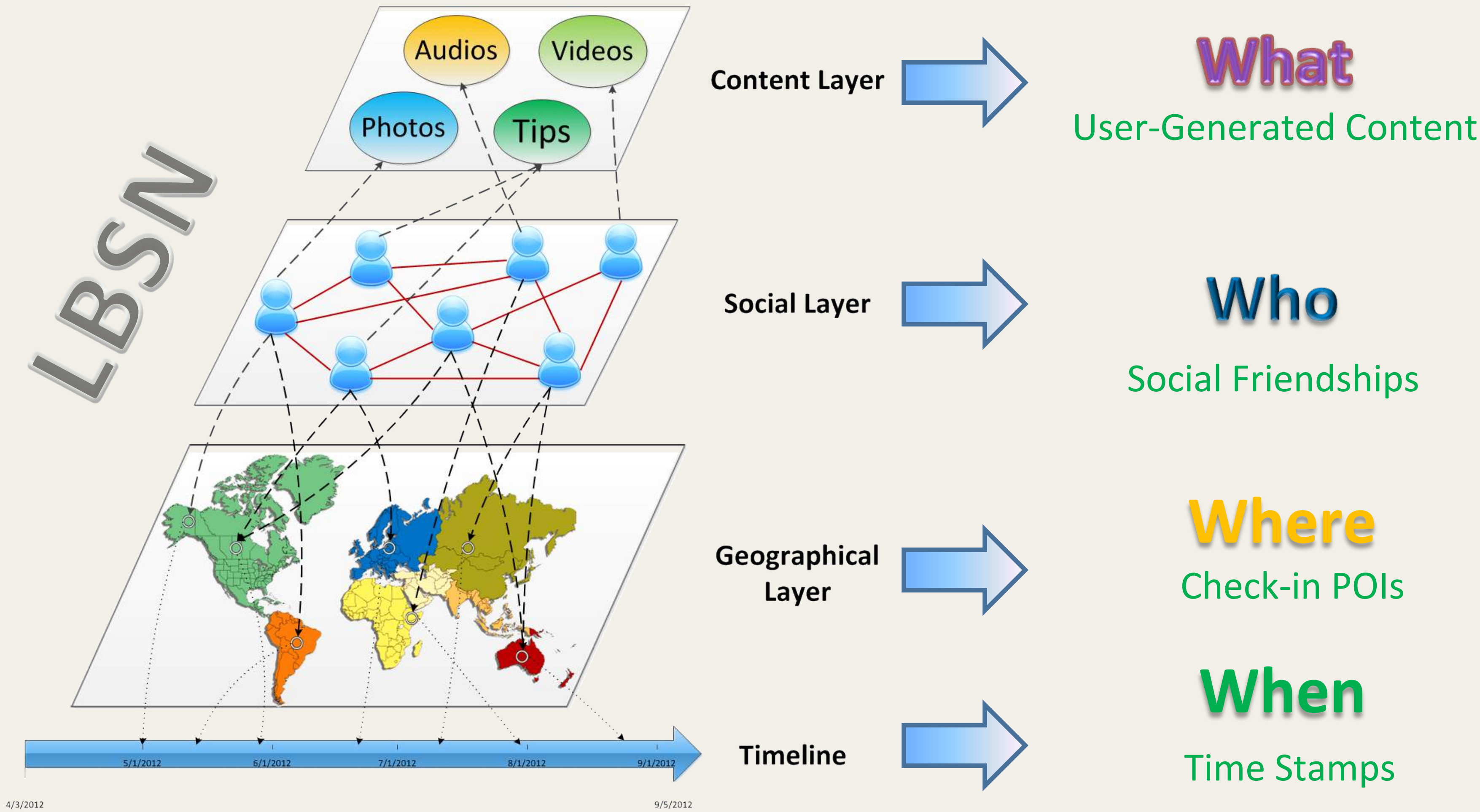
- Introduction
- Content Recommendation
- Location Recommendation
- Future Work

Location-based Social Networks [Gao et al., 2014,2015]

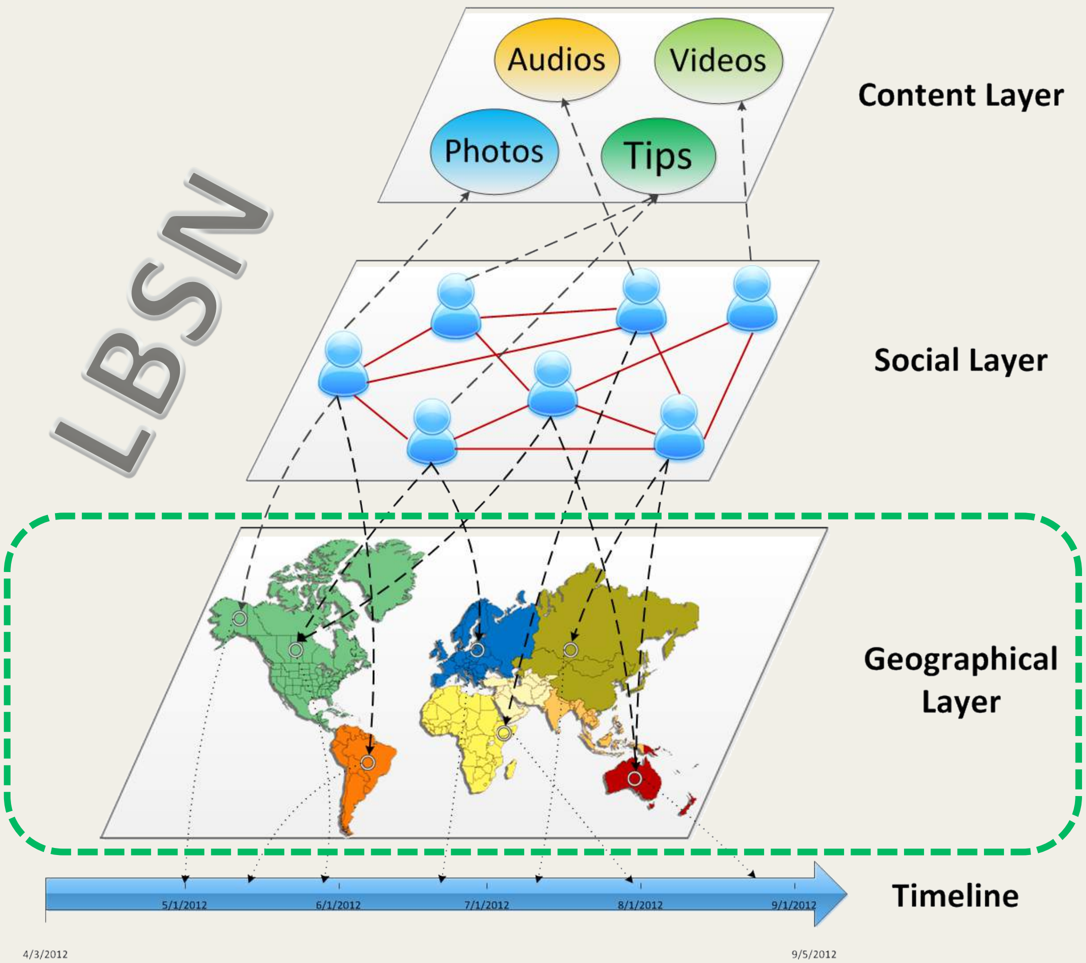
- Location-Based Social Networking Sites
  - Foursquare, Facebook Places, Yelp



Information in LBSNs



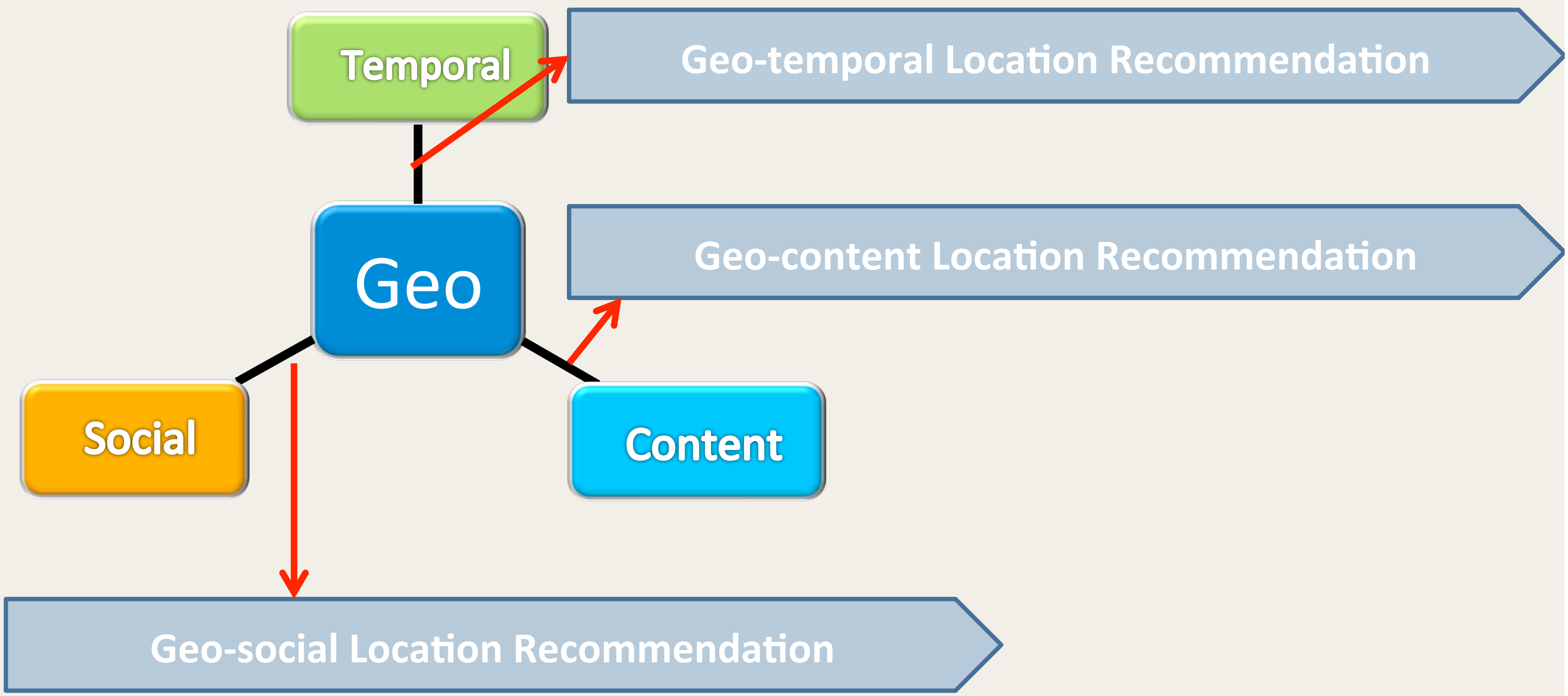
Challenges for Location Recommendation



- Geographical Properties of Social Connections
  - Geographical Distance
  - Social Connections
- Temporal Cyclic Patterns of Geographical Check-ins
  - Going to a restaurant around noon
  - Watching movie in a theater during the weekend
- Content information could be important

Categorization of Location Recommendation

- Location recommender systems can be divided into three groups according to the information used



Location Recommendation in Social Media

Location Recommendation

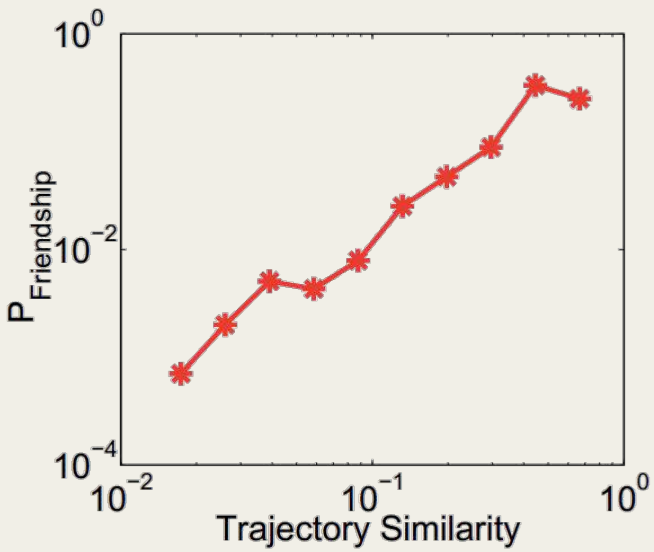
Geo-social Location Recommendation

Geo-temporal Location Recommendation

Geo-content Location Recommendation

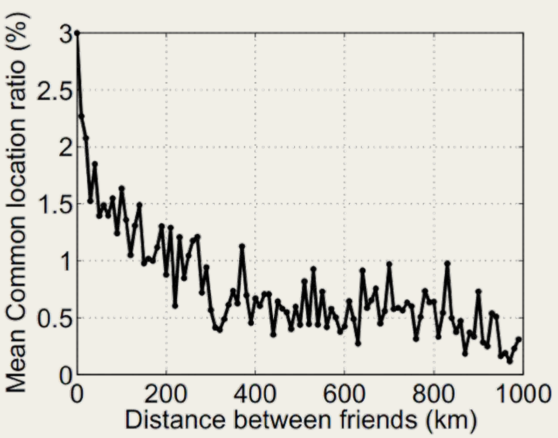
Geographical Properties of Social Connections

- There is a strong correlation between friendship and trajectory similarity in LBSNs



[Cho et al., 2011]

- Nearby friends have a much higher probability to share common locations



[Mao et al., 2010]

Friend-based Methods [Mao et al., 2010]

- Friend-based Collaborative Filtering: FCF

$$\widehat{r}_{i,j} = \frac{\sum_{u_k \in U_i'} r_{k,j} w_{i,k}}{\sum_{u_k \in U_i'} w_{i,k}}$$

Trajectory Similarity

- Geo-Measured FCF: GM-FCF
  – Assuming a power-law relation between trajectory similarity y and geographical distance x

$$y = \alpha x^{\beta}$$

  – Similarity is computed as

$$w_{i,k} = \frac{y(x = d(u_i, u_k), \alpha, \beta)}{\sum_{u_k \in U_i} y(x = d(u_i, u_k), \alpha, \beta)}$$

Preference and Friend based methods

- A fusion model: USG **[Mao et al., 2011]**
  – The probability score of i-th user at j-th location is

$$S_{i,j} = (1-\alpha-\beta)S^{u}_{i,j} + \alpha S^{s}_{i,j} + \beta S^{g}_{i,j}$$

User preference: User-oriented CF

Social Influence: FCF

Geographical Influence

- A Social-Historical Model: SHM **[Gao et al., 2012a]**
  – Users' historical information is modeled by Hierarchical PitmanYor process

Language Modeling		LBSN Modeling	
Corpus		Check-in Collection	
Document		Individual's Check-ins	
Document Structure	Paragraph	Check-in Structure	Monthly Check-in Sequence
	Sentence		Weekly Check-in Sequence
	Phrase		Daily Check-in Sequence
	Word		Check-in Location

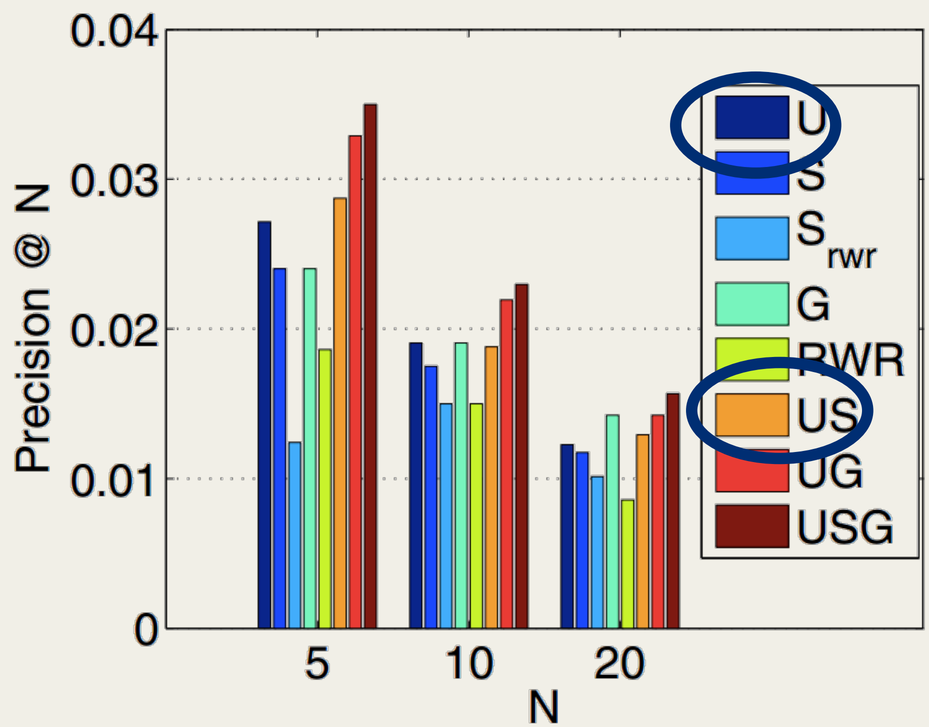
$$P^{SH}_{i,j} = \alpha P_{i,j} + (1-\alpha)\sum_{u_k \in N_i} w_{i,k} P_{k,j}$$

User preference

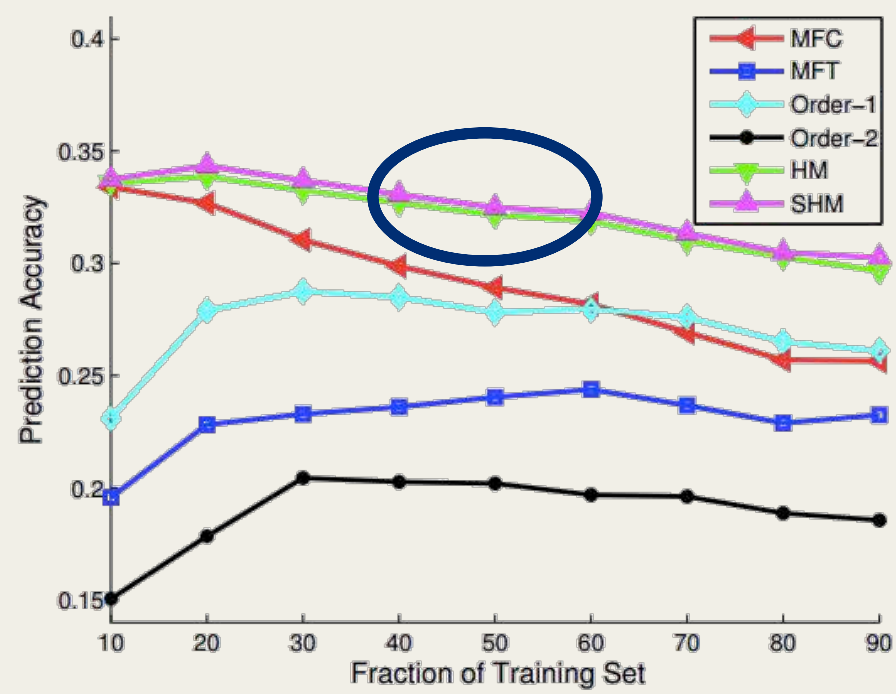
Social Influence

Some Observations for Geo-social Location Recommendation

- Social information can consistently improve the recommendation performance, however, the improvement is very limited



[Mao et al., 2011]



[Gao et al., 2012a]

Geo-social Circles [Gao et al., 2012b]

- Friends with long distance share a small number of commonly visited locations
- Non-friends with short distance share a large number of commonly visited locations
- Users are segmented into four geo-social circles

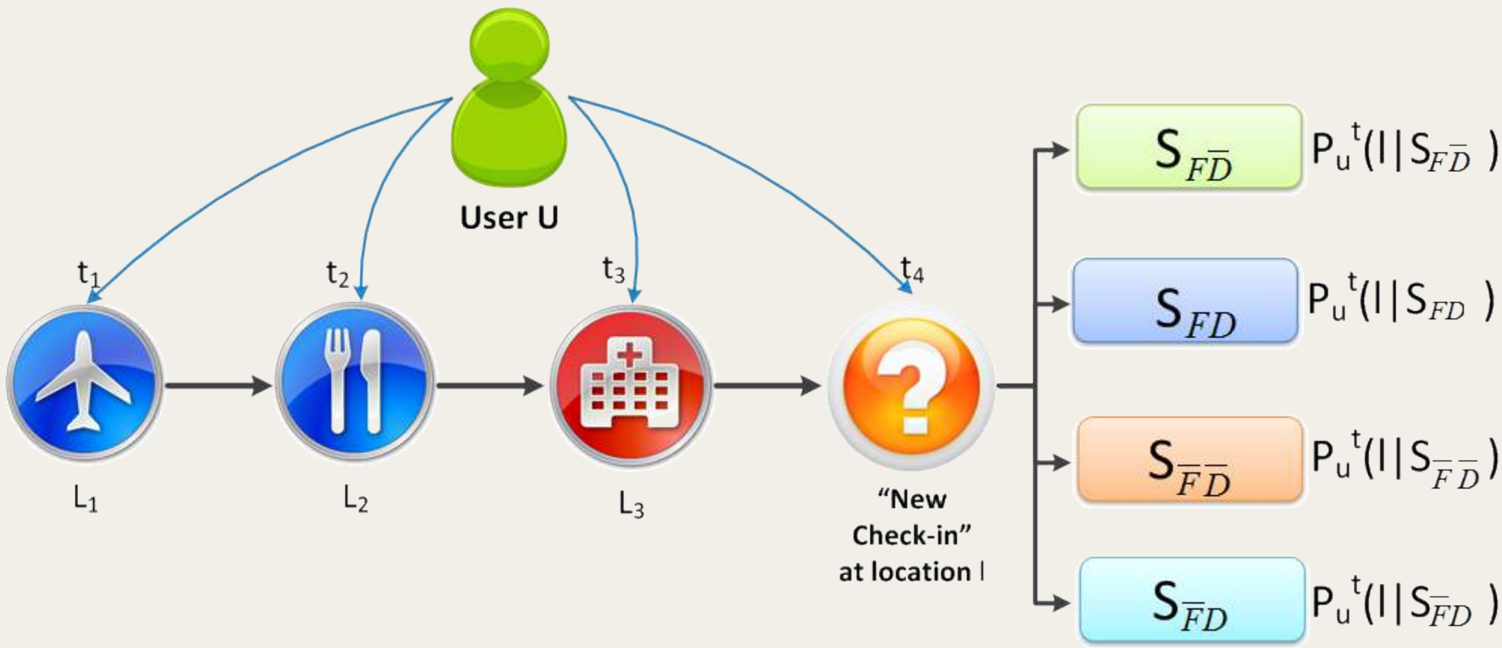
Geo-Social Circles

	F	$\bar{F}$
$\bar{D}$	$S_{F\bar{D}}$: Local Friends	$S_{\bar{F}\bar{D}}$: Local Non-friends
D	S_{FD}: Distant Friends	$S_{\bar{F}D}$: Distant Non-friends

A Cold-start Location Recommendation Framework

- A framework is proposed to address cold-start problem in location recommendation based on geo-social circles

$$P_u^t(l) = \Phi_1 P_u^t(l|S_{\bar{F}\bar{D}}) + \Phi_2 P_u^t(l|S_{F\bar{D}}) + \Phi_3 P_u^t(l|S_{FD}) + \Phi_4 P_u^t(l|S_{\bar{F}D}).$$



Observations about Geo-social Circles

- Local friends are more important than distant friends
- Distance friends contain more additional information than local friends when combining with local non-friends
- These four geo-social circles contain complementary information although their contributions differ

Methods	Top-1	Top-2	Top-3
$S_{F\bar{D}}$	6.51%	8.31%	9.32%
S_{FD}	3.65%	4.75%	5.34%
$S_{\bar{F}\bar{D}}$	18.37%	24.10%	27.34%
$S_{\bar{F}\bar{D}} \cup S_{F\bar{D}}$	18.62%	24.44%	27.79%
$S_{\bar{F}\bar{D}} \cup S_{FD}$	19.01%	24.95%	28.35%
$S_{F\bar{D}} \cup S_{FD}$	8.33%	10.79%	12.23%
$S_{\bar{F}\bar{D}} \cup S_{F\bar{D}} \cup S_{FD}$	**19.21%**	**25.19%**	**28.69%**

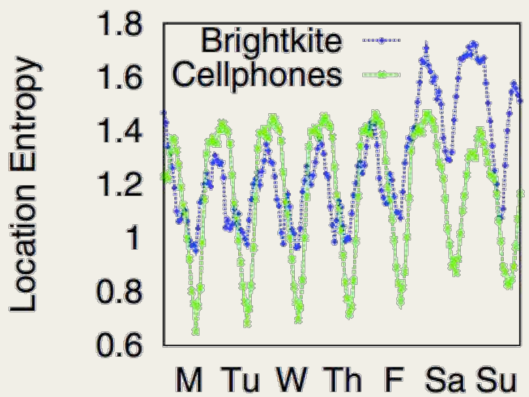
Location Recommendation in Social Media

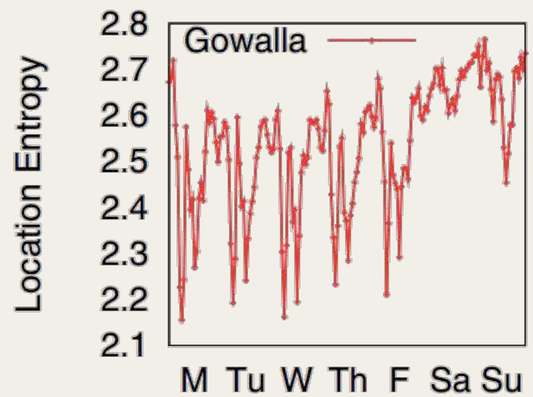
Location Recommendation

- Geo-social Location Recommendation
- Geo-temporal Location Recommendation
- Geo-content Location Recommendation

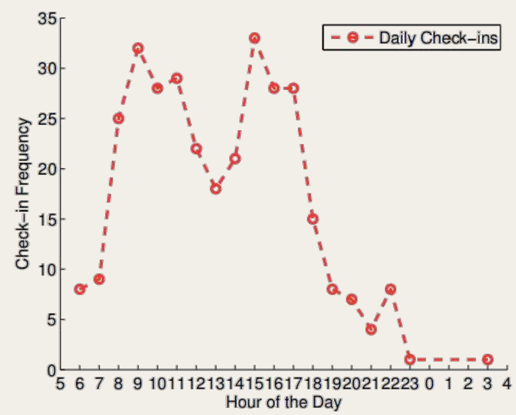
Why Temporal Information Matters?

- Human movement exhibits strong temporal cyclic patterns
  - Days of the week patterns **[Cho et al., 2011]**





  - Hours of the day patterns [**Gao et al., 2013a**]



Location Recommendation with Time Preference: UT [Yuan et al., 2013]

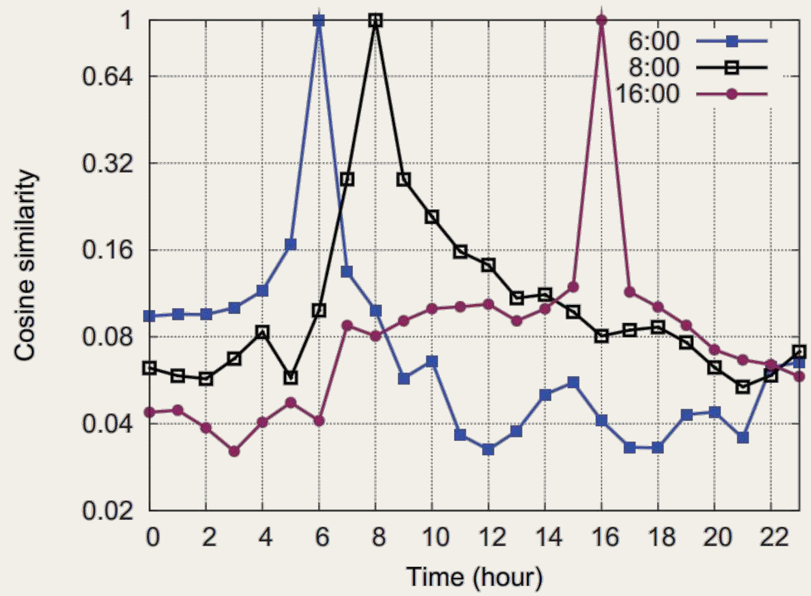
- Splitting data into 24 slots based on hours
  – Nov. 6 2012, 10:30 → 10
- Introducing time dimension into user-location matrix c
  – $c_{u,l} \rightarrow c_{u,t,l}$
- Leveraging time factor when
  – Computing the similarities between users over time

$$w_{u,v}^{(t)} = \frac{\sum_t \sum_l c_{u,t,l} c_{v,t,l}}{\sqrt{\sum_t \sum_l c_{u,t,l}^2} \sqrt{\sum_t \sum_l c_{v,t,l}^2}}$$

  – Making predictions $\hat{c}_{u,t,l} = \frac{\sum_v w_{u,v}^{(t)} c_{v,t,l}}{\sum_v w_{u,v}^{(t)}}$

Enhancing UT by Smoothing

- Data in each slot becomes even sparser after splitting
- Check-in behaviors of users at different time are correlated
- Smoothing $c_{u,t,l}$ based on the similarity between different time slots



$$\widetilde{c}_{u,t,l} = \sum_{t'=1}^{T} \frac{\rho_{t,t'}}{\sum_{t''=1}^{T} \rho_{t,t''}} c_{u,t',l}$$

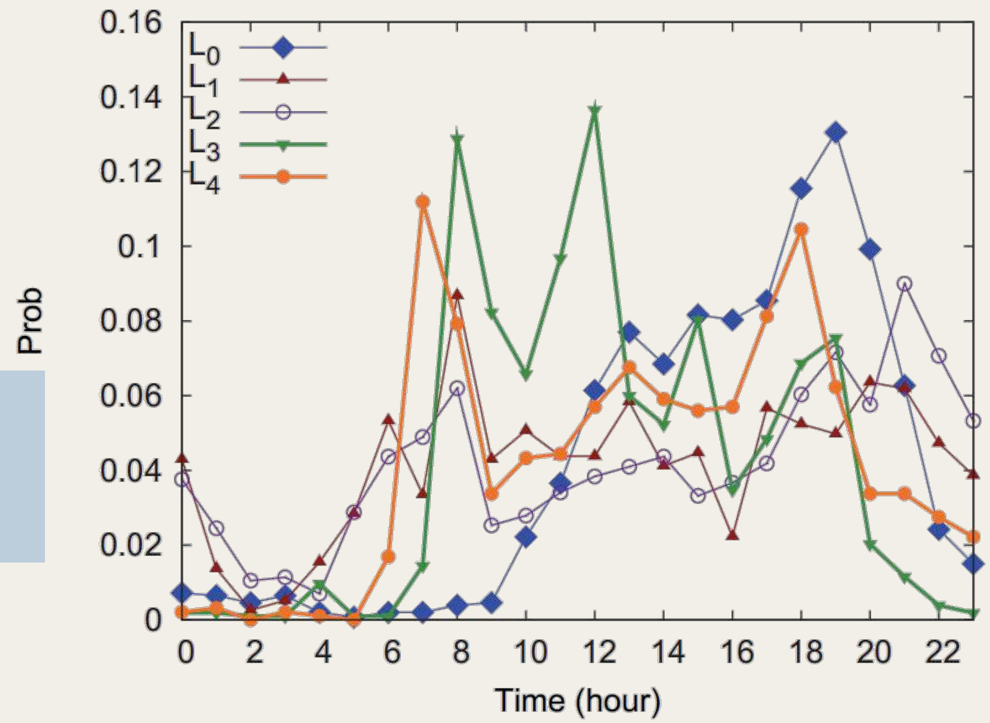
Enhancing UT by Location Popularity

- The popularity of a location varies over time
  - A restaurant is more popular around noon and evening
- Location popularity is calculated as

$$P_t(l) = \beta \frac{|CI_{l,t}|}{\sum_{l' \in L} |CI_{l',t}|} + (1-\beta) \frac{|CI_l|}{\sum_{l' \in L} |CI_{l'}|}$$

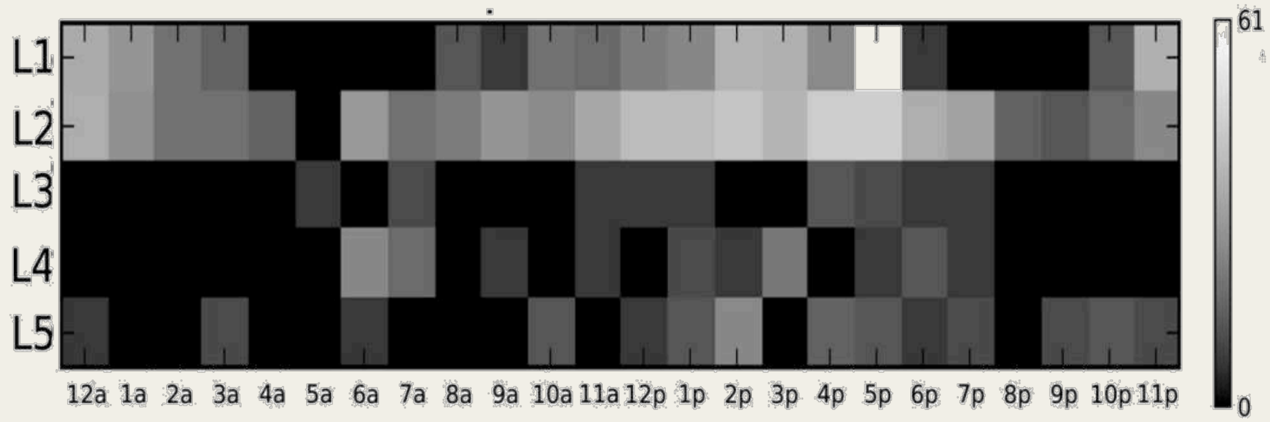
Number of Check-ins at *l* at time *t*

Number of Check-ins at *l*



Location Recommendation with Temporal Effects [Gao et al., 2013]

- One user's daily check-in activity w.r.t. his top 5 frequently visited locations



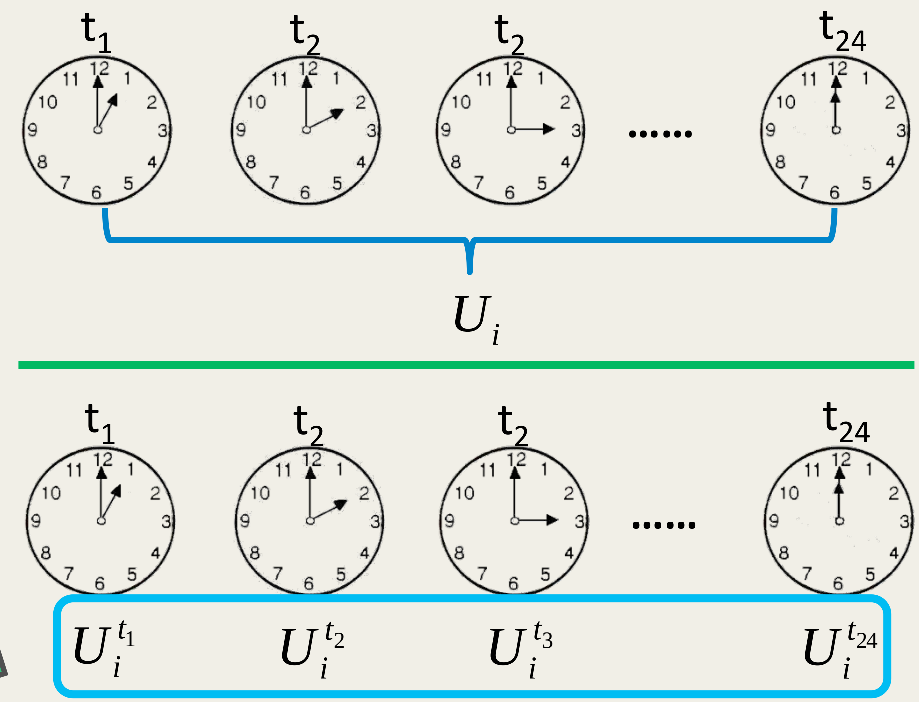
- Temporal Non-uniformness
  - A user presents different check-in preferences at different hours of the day
- Temporal Consecutiveness
  - A user presents similar check-in preferences at nearby hours of the day

Modeling Temporal Non-uniformness

- A user presents different check-in preferences at different hours of a day

$$\min_{U_i \ge 0, L_j \ge 0} \sum_{i}^{m} \sum_{j}^{n} Y_{i,j} (C_{i,j} - U_i L_j^T)^2$$

$$\min_{U_i \ge 0, L_j \ge 0} \boxed{\sum_{t=1}^{24}} \sum_{i}^{m} \sum_{j}^{n} Y_{i,j}^t (C_{i,j}^t - \boxed{U_i^t} L_j^T)^2$$



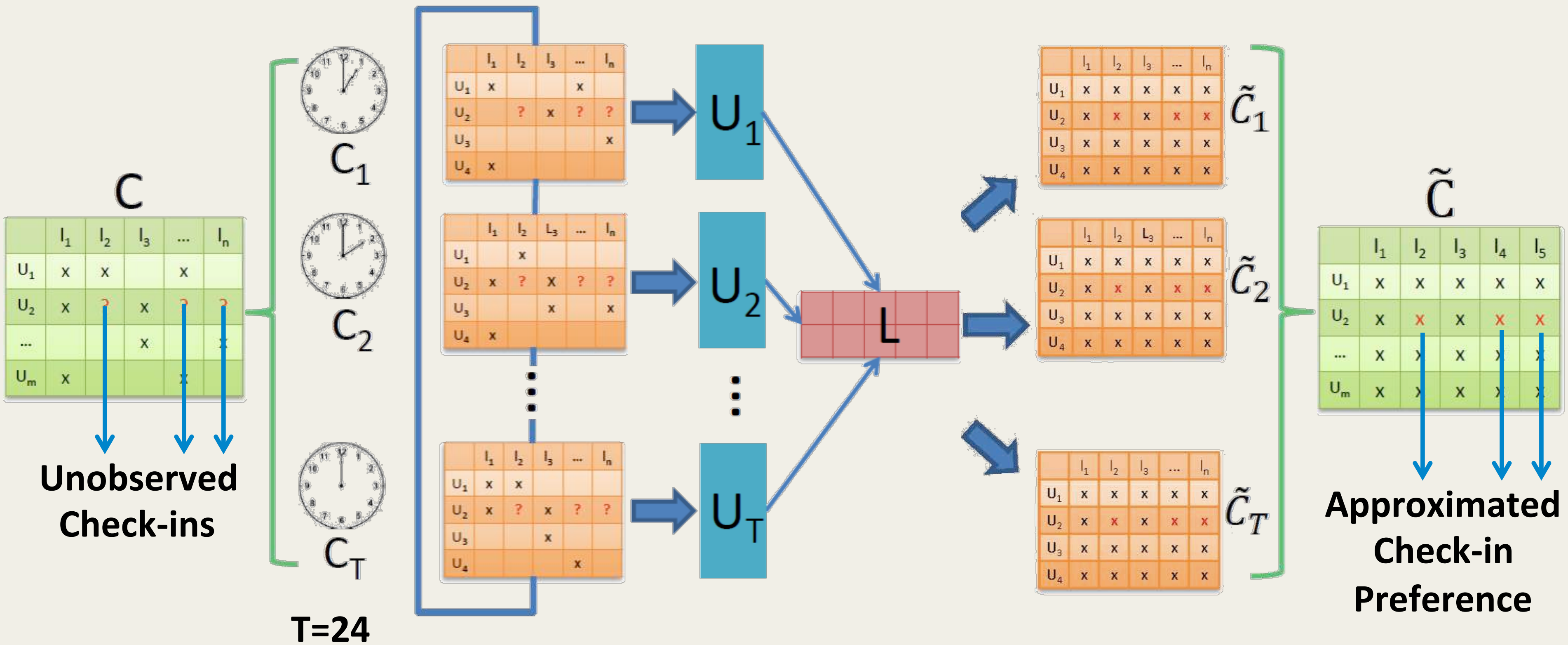
Modeling Temporal Consecutiveness

- A user presents similar check-in preferences at nearby hour of the day

$$\min_{U \geq 0} \sum_{t=1}^{T} \sum_{i=1}^{m} \boxed{\psi_i(t, t-1)} \boxed{\left\| U_t(i,:) - U_{t-1}(i,:) \right\|_F^2}$$

$$\psi_i(t, t-1) = \frac{C_t(i,:) \cdot C_{t-1}(i,:)}{\sqrt{\sum_j C_t^2(i,:)} \sqrt{\sum_j C_{t-1}^2(i,:)}}$$

Framework of Location Recommendation with Temporal Effects



Location Recommendation in Social Media

Location Recommendation

Geo-social Location Recommendation

Geo-temporal Location Recommendation

Geo-content Location Recommendation

Content in LBSNs

- Content in LBSNs is usually available
  - Tags, tips or comments

- Content contains semantic words that reflect a user's interested topics and the location property
  - ``Chinese'' and ``Spicy''
- Content can reflect users' preferences
  - ``all great"

Why Sentiment in Content is Important?

- Ratings in traditional recommendation can capture user preferences
  - Like/dislike, voting scores from 1 to 5

- Check-in behavior represents users' habitual behavior and may not be sufficient to reflect users' preferences
  - High check-in frequencies may represent positive opinions
  - Fewer checked locations are not necessarily less favored

- Sentiment extracted from content contains more precise information about a user's preference on a location
  - In addition to positive feedback, there could also be negative feedback from content

Sentiment-enhanced Location Recommendation [Yang et al., 2013]

- Extracting check-in preferences from check-in data
- Extracting sentiment preferences from content
- Combining check-in preferences and sentiment preferences
- Performing traditional CF based on the combined preferences

Preference Extraction from Check-ins

- Check-in frequencies can reflect users' preferences
  - Users prefer those locations with high check-in frequencies

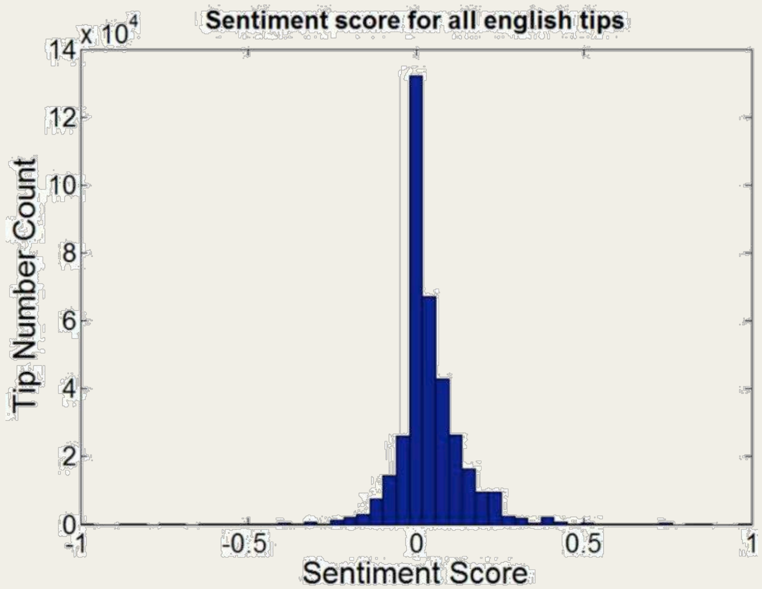
- Mapping frequencies to five-point preferences
  - Check-in frequencies follow the power law distribution

Frequency	Preference Scores
1	2 (Fair)
2	3 (Good)
3	4 (Very Good)
>=4	5 (Excellent)

- Constructing a check-in preference matrix P_c

Preference Extraction from Content

- Sentiment extracted from content could reflect user's preference on a location
- Mapping sentiment scores to five-point preferences
  - Sentiment scores are highly centralized around 0
  - A slight bias towards positive sentiment



Sentiment Scores	Preference Scores
[-1,-0.05]	1
(0.05,-0.01]	2
(-0.01,0.01)	3
[0.01,0.05)	4
[0.01,1]	5

- Constructing a sentiment preference matrix P_s

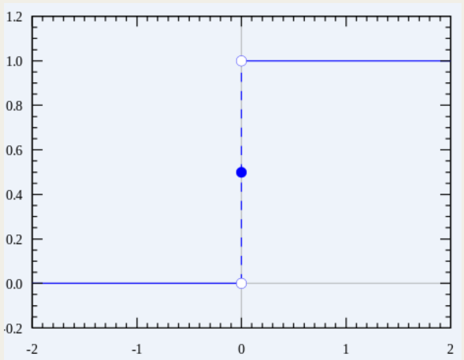
Sentiment-enhanced Location Recommendation [Yang et al., 2013]

- Combining the check-in preference matrix and the sentiment preference matrix
  - Sentiment preference has a bigger impact for one-time check-in locations
  - Sentiment preference has some impact for multi-time check-in locations

$$P_{final} = P_c + sgn(P_c - P_s) \cdot H(|P_c - P_s| - 2)$$

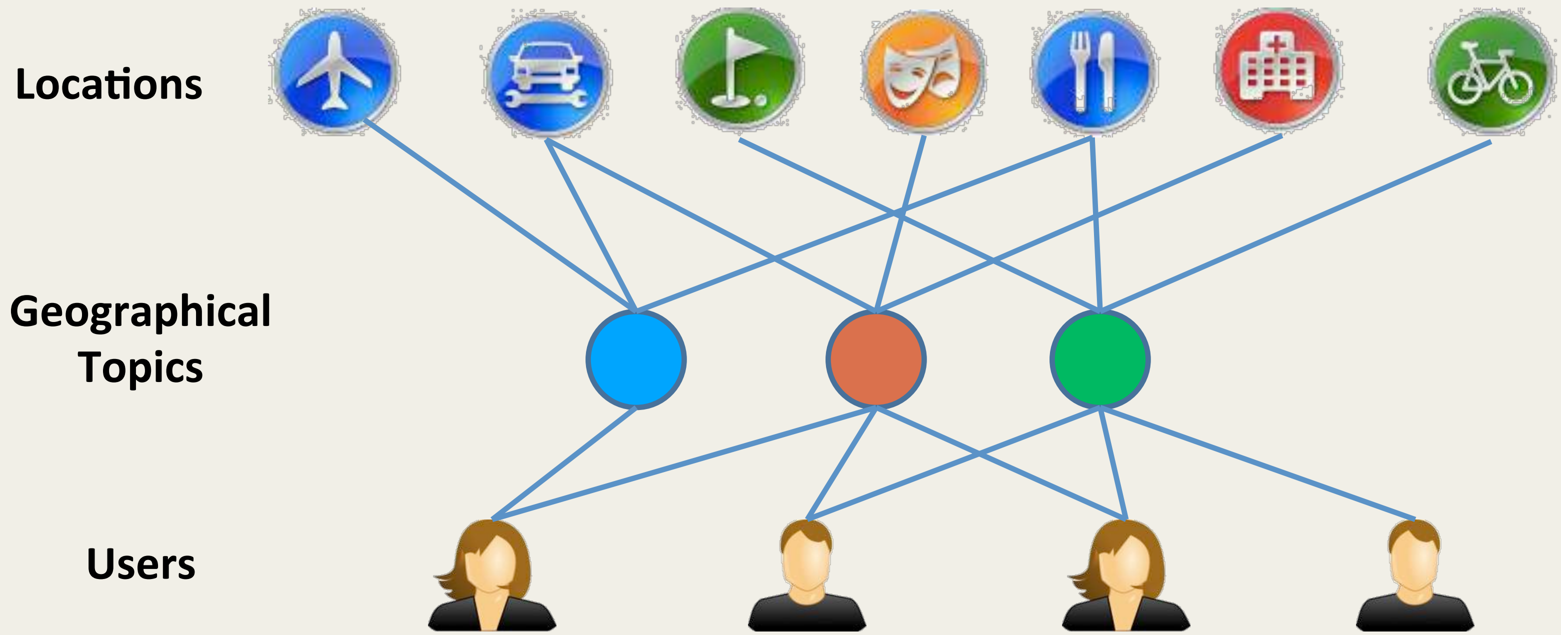
Sign Function

Heaviside Step Function



Geographical Topics from Content in LBSNs

- Geographical topics are discovered from LBSNs **[Yin et al., 2011]**
  – Assigning semantic topics to locations
  – Reflecting users' interests
  – Connecting users and locations in the semantic level



Topic-aware Location Recommendation [Liu and Xiong, 2013]

- Building an aggregated LDA model to discover geographical topics
  - User interest topic distribution θ_i
  - Location topic distribution π_j
- Defining topic and location influence index

$$TL_{ij} = \alpha(1 - \underbrace{D_{JS}}_{\text{Jensen-Shannon Divergence}}(\theta_i, \pi_j)) + (1-\alpha)\underbrace{P_j}_{\text{Location Popularity}}$$

- Modeling users check-in behaviors as

$$c_{ij} = TL_{ij} U_i^T C_j$$

References

[**Gao et al., 2014**] H. Gao, J. Tang, H. Liu. Personalized Location Recommendation on Location-based Social Networks, pp, 399—400, 2014.

[**Scellato et al., 2011**]S. Scellato, A. Noulas, R. Lambiotte, C. Mascolo. Socio-Spatial Properties of Online Location-Based Social Networks. In ICWSM, pp, 329—336, 2011.

[**Cho et al., 2011**] E. Cho, S. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp, 1082—1090, 2011.

[**Mao et al., 2010**] Y. Mao, P. Yin, and W. Lee. Location recommendation for location-based social networks. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp, 458—461, 2010.

[**Mao et al., 2011**]Y. Mao, P. Yin, W. Lee, and D. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp, 325—334, 2011.

[**Gao et al., 2012a**] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In ICWSM, 2012.

[**Gao et al., 2012b**] H. Gao, J. Tang, and H. Liu. gSCorr: modeling geo-social correlations for new check-ins on location-based social networks. In Proceedings of the 21st ACM international conference on Information and knowledge management, pp, 1582—1586, 2012.

References

[**Yuan et al., 2013**] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Thalmann. Time-aware Point-of-interest Recommendation. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp, 363—372, 2013.

[**Gao et al., 2013b**] H. Gao, J. Tang, H. Xia, and H. Liu. Exploring temporal effects for location recommendation on location-based social networks. In Proceedings of the 7th ACM conference on Recommender systems, pp, 93—100, 2013.

[**Yin et al., 2011**] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. Proceedings of the 20th international conference on World wide web, pp, 247—256, 2011.

[**Yang et al., 2013**] D. Yang, D. Zhang, Z. Yu, and Z. Wang. sentiment-enhanced personalized location recommendation system. In Proceedings of the 24th ACM Conference on Hypertext and Social Media, pp, 119—128, 2013.

[**Liu and Xiong, 2013**] B. Liu, and H. Xiong. Point-of-Interest Recommendation in Location Based Social Networks with Topic and Location Awareness. SIAM Conference on Data Mining, pp, 396—404, 2013.

[**Gao and Liu, 2015**] H. Gao, and H. Liu. Mining Human Mobility in Location-based Social Networks, Synthesis Lectures on Data Mining and Knowledge Discovery, pp, 1—115, 2015.

Outline

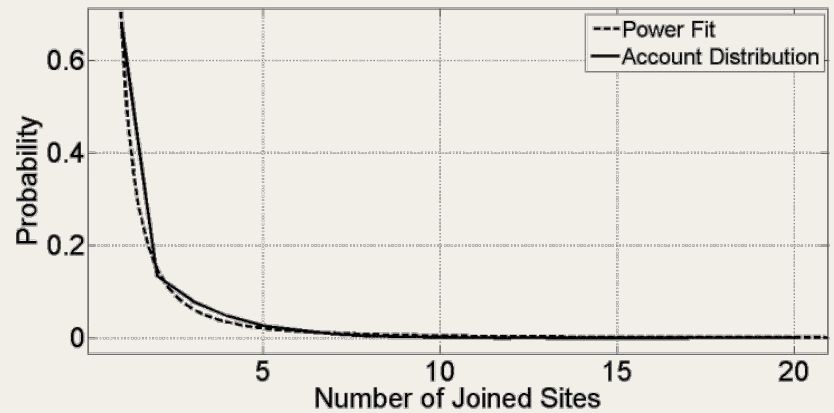
- Introduction
- Content Recommendation
- Location Recommendation
- Future Work

Recommendation with Cross-Media Data

- Users usually join multiple social media sites

[Zafarani and Liu, 2014]

  – More than 97% of users have joined at most 5 sites
  – Users exist on as many as 16 sites



- A new user on one site might have existed on other sites for a long time
  – Cross media data can mitigate data sparsity problem
  – Cross media data can reduce cold-start users

Deep Learning in Recommendation

- Deep learning has been proven to be effective in various domains
  - Pattern recognition and natural language processing
- Recently deep convolutional neural networks is used to predict latent factors from music audio for music recommendation **[VanDeOord et al., 2013]**
  - A content-based method without data sparsity problem in collaborative filtering
  - Viable for recommending new and unpopular music
- How to apply deep learning with rich social media data is still an open issue

Privacy-preserving Recommendation

- Recommender systems in social media may utilize sensitive information from users to produce better recommendations
  - Users' locations in location-aware content recommendation
  - Social networks in social recommendation
  - Check-in data in location recommendation
- New privacy threats are introduced by recommender systems in social media **[Jeckmans et al., 2013]**
  - The privacy of social relations
  - The privacy of their locations

References

[**Zafarani and Liu, 2014**] R. Zafarani, and H. Liu. Users Joining Multiple Sites: Distributions and Patterns. In International AAAI Conference on Weblogs and Social Media, 2014.

[**VanDeOord et al., 2013**] A. Van Den Oord, S. Dieleman, and B. Schrauwen. Deep content-based music recommendation. In Advances in Neural Information Processing Systems, pp, 2643—2651, 2013.

[**Jeckmans et al., 2013**] A. Jeckmans, M. Beye, Z. Erkin, P. Hartel, R. Lagendijk and Q. Tang. Privacy in Recommender Systems. In Social Media Retrieval, pp, 263—281, 2013.

[**Tang et al., 2014**] J. Tang, J. Tang, and H. Liu. Recommendation in Social Media: Recent Advances and New Frontiers. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp, 1977-1977, 2014.

[**Gao et al., 2014**] H. Gao, J. Tang, H. Liu. Personalized Location Recommendation on Location-based Social Networks, pp, 399—400, 2014.

[**Gao and Liu, 2015**] H. Gao, and H. Liu. Mining Human Mobility in Location-based Social Networks, Synthesis Lectures on Data Mining and Knowledge Discovery, pp, 1—115, 2015.

Acknowledgements

- Projects are partially supported by **National Science Foundation, Army Research Office** and **The Office of Naval Research**

- Members of Data Mining and Machine Learning Lab at ASU provided valuable feedback and suggestions