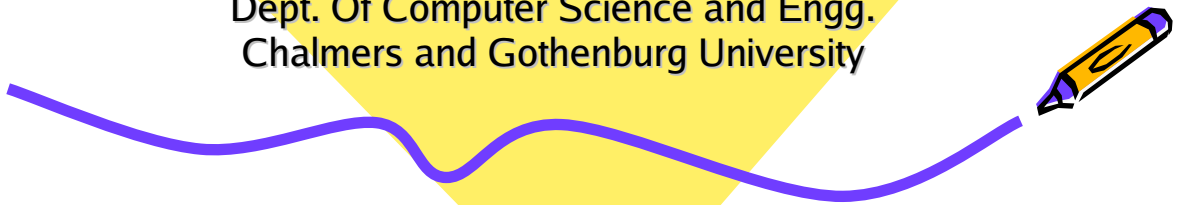


The Wisdom of Crowds (Who Surf)

Devdatt Dubhashi
Dept. Of Computer Science and Engg.
Chalmers and Gothenburg University



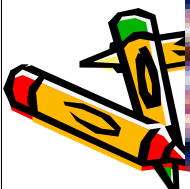
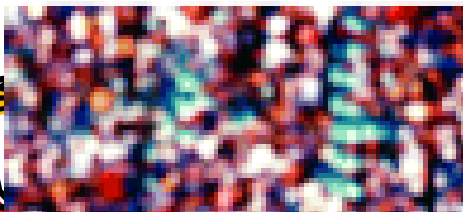
The Wisdom of Crowds

A NEW YORK TIMES BUSINESS BESTSELLER
"An extraordinary and thought-provoking... The Wisdom of Crowds..."
—The Boston Globe

**THE WISDOM
OF CROWDS**

**JAMES
SUROWIECKI**

WITH A NEW INTRODUCTION BY THE AUTHOR



- Why the Many are smarter than the Few
- Crowds better than experts
- Discovery
- Coordination
- Cooperation



Francis Galton visits a Country Fair (1906)



- Crowd guesses what the Ox would weigh
- Crowd's average guess: 1,197 lb
- Correct weight: 1,198 lb!



Wisdom of Crowds on the Web



- Exponentially growing
- Also quite accurate: a recent comparison with Encyclopaedia Britannica



WIKIPEDIA
The Free Encyclopedia



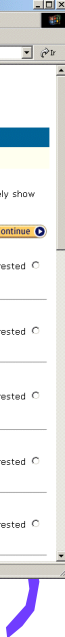
Wisdom of Crowds on the Web: Google

- Link analysis
- Basic idea is that more people linking to a page is more votes for it
- Subtle re-inforced version of the "wisdom of crowds"



Mining the Internet

- Collaborative Filtering
- Simultaneous categorization of people and products based on mathematical models.



Amazon.co.uk: Your Recommendations - Microsoft Internet Explorer

Address: http://www.amazon.co.uk/your-recommendations/instants-recof/books/ubatch/none/0/01/0c/ef-epd_r_batch_14/026-0668321-0070807

Amazon.co.uk
MasterCard
Apply now
more info

amazon.co.uk

VIEW BASKET | WISH LIST | YOUR ACCOUNT | HELP

WELCOME | A'S STORE | BOOKS | ELECTRONICS & PHOTO | MUSIC | DVD | VIDEO | SOFTWARE | PC & VIDEO GAMES | HOME & GARDEN | TOYS & KIDS | TRAVEL

YOUR FAVOURITE STORES | YOUR RECOMMENDATIONS | THE PAGE YOU MADE | NEW FOR YOU

Hello J.C. Dursteler Lopez, we have recommendations for you (if you're not J.C. Dursteler Lopez, [click here](#)). Here's your [Like For You](#)™ recommendations

Your Recommendations > Books

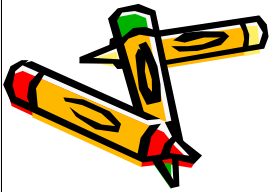
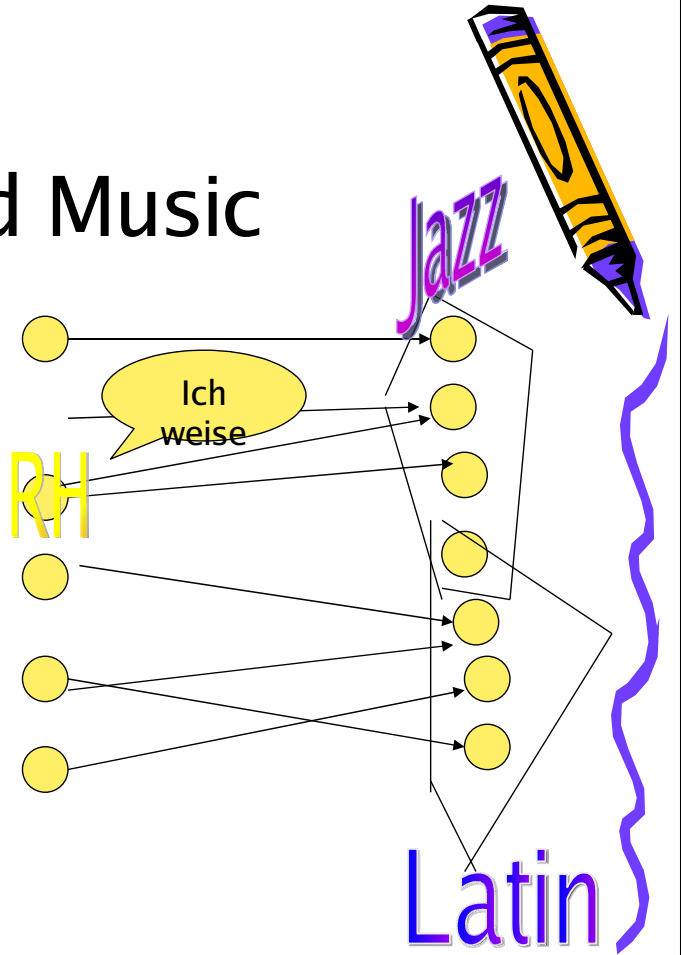
Already own any of these titles? Know you won't like one? Refine your recommendations and we'll immediately show you new choices!

To save your choices and get new recommendations, click [Save & Continue](#)

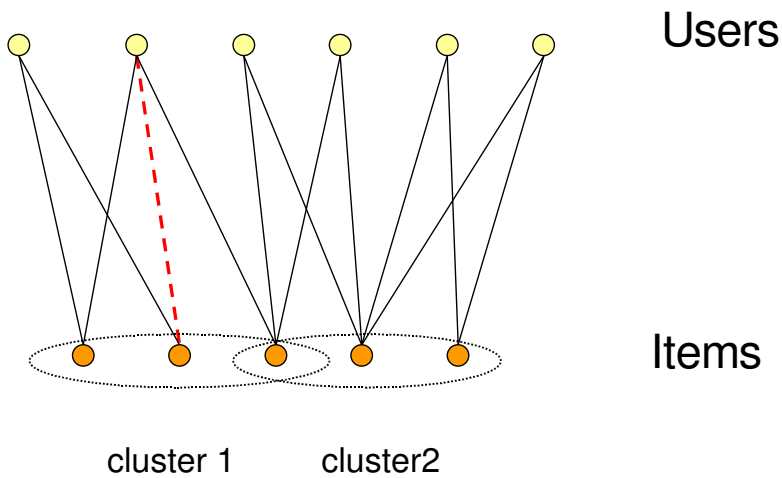
Item	Title	Author	No Opinion	I own it	Not interested
1.	Content Critical: Gaining Competitive Advantage Through High-Quality Web Content	By Gerry McGovern, Rob Norton	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2.	The Elements of User Experience	By Jesse James Garrett	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3.	The Design of Sites: Principles, Processes and Patterns for Crafting a Customer-Centered Web Experience	By Douglas K-Van Duijne, et al	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4.	Information Architecture: Blueprints for the Web	By Christina Wodtke	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5.	Designing with Web Standards	By Jeffrey Zeldman	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

People and Music

- People and music they buy often shows a distinct bi-clustering
- Mathematical Mixture models try to capture this structure



Collaborative filtering

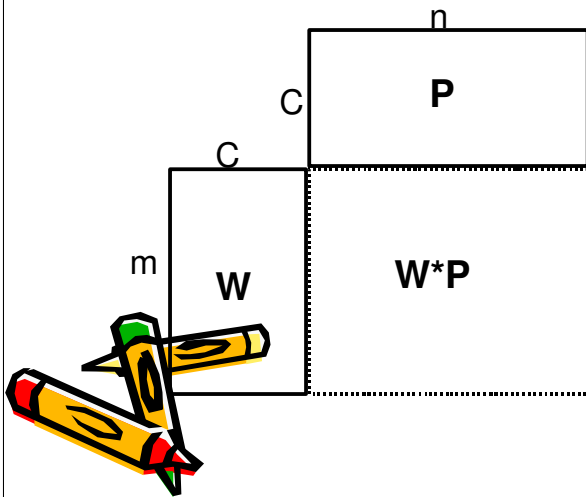


Mixture model

$$\forall u : \text{preference } p(u, c) \quad \left(\sum_c p(u, c) = 1 \right)$$

$$\forall a : \text{weight } w(a, c) \quad \left(\sum_a w(a, c) = 1 \right)$$

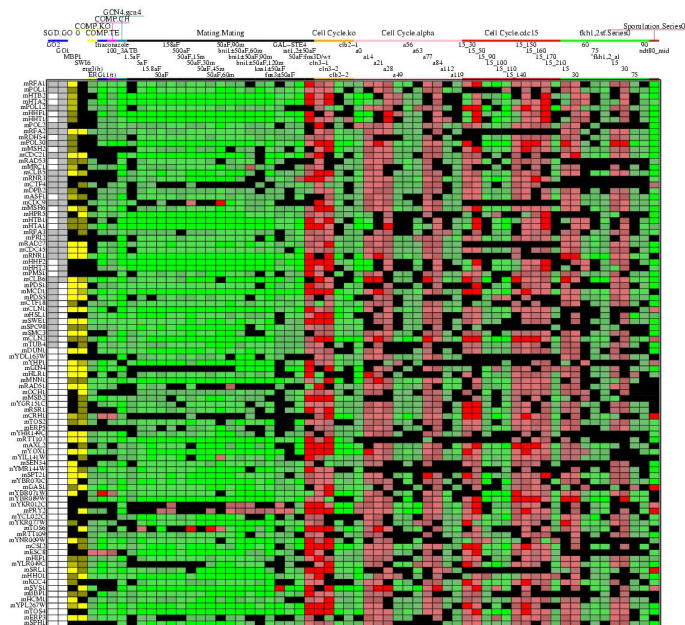
$$\Pr(u \text{ selects } a) = \sum_c p(u, c)w(a, c)$$



number of users: n
 number of items: m
 number of clusters: C

Biclustering Genes

Genes

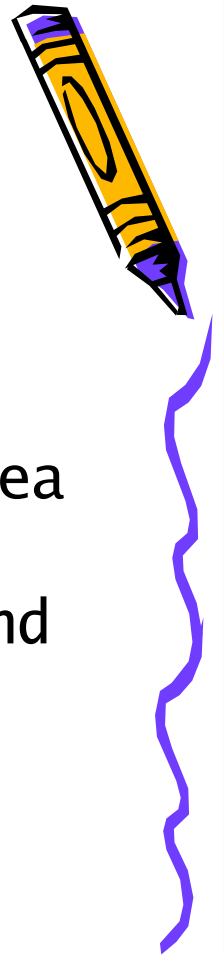


Disease exps

GO1 - DNA metabolism (GO:0006259)
 GO2 - mitotic cell cycle (GO:0000278)

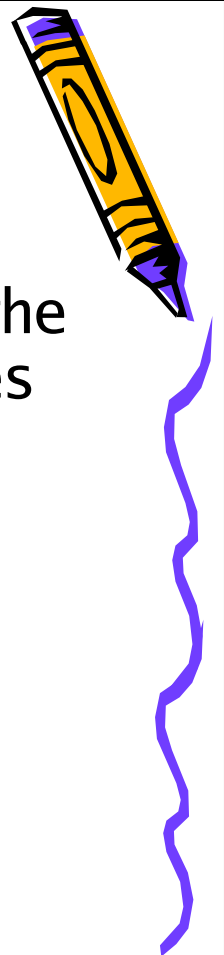
Bi-clustering algorithms

- Exploit clustering structure simultaneously on both sides.
- "Biclustering is a relatively young area ... it has great potential to make significant contributions to biology and **other fields**." [Tanay, Sharan and Shamir, 2006]



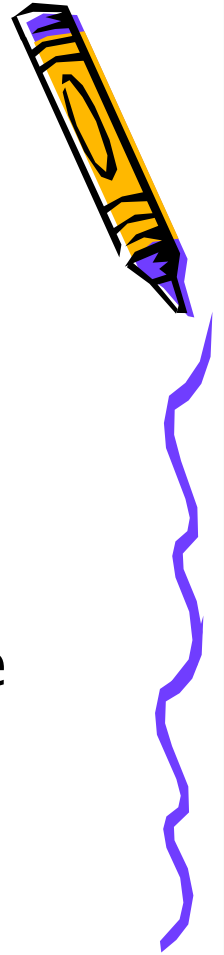
Collaborative Filtering in Mixture Models

- Kleinberg and Sandler (2004) gave the first rigorous analysis and guarantees on algorithms for the mixture model
- Uses LP and spectral methods
- We wanted to explore light-weight practical methods.



A Porfolio of Iterative Bi-clustering Algorithms

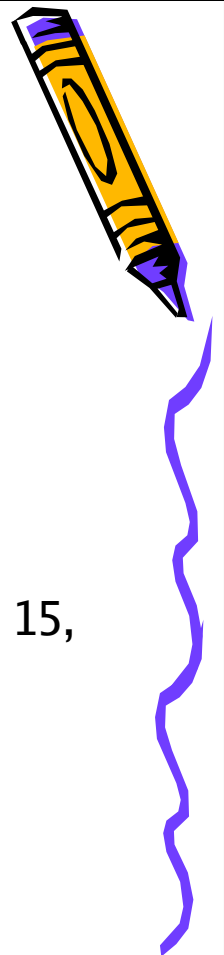
- Start with an initial soft clustering on one side.
- (**HITS**) Iteratively use the current clustering on one side to refine the clustering on the other side
- Parameters: no. of iterations, update weighting.



Tests on generated data

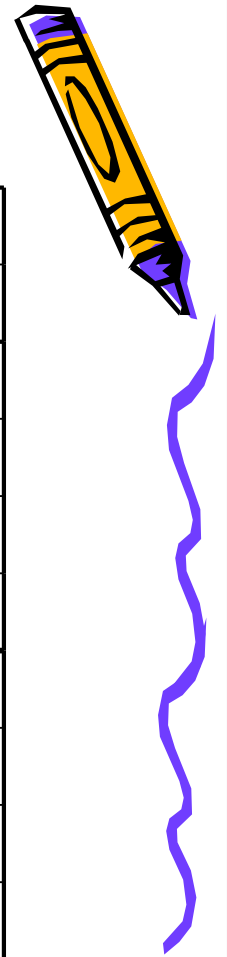
Planted partition Model

- Disjoint model
- Intermediate model
- Completely mixed model
- Generated:
 - 1000 users
 - 300 items
 - 10 clusters
 - sample size: 10, 15, 20
- $\Theta \in [0,1]$



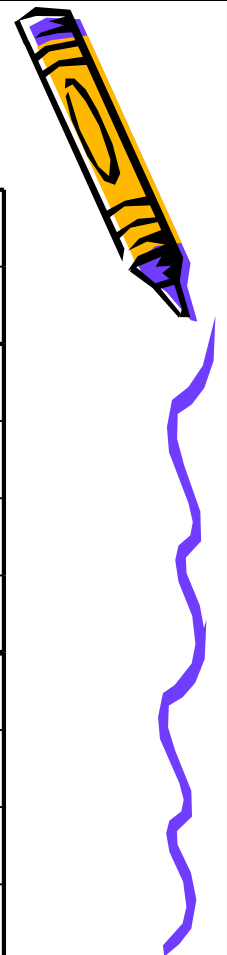
Results for the disjoint model

		Sample 10		Sample 20	
Θ	Method	F10	F20	F10	F20
0	1	4.7	14.9	6.3	16.1
	2	4.2	14.5	6.0	15.8
	3	4.0	14.3	5.9	15.8
	4	4.3	14.6	6.1	15.9
0.4	1	8.7	19.0	8.7	19.4
	2	3.7	13.9	6.2	16.1
	3	3.7	14.0	6.2	16.2
	4	3.7	14.0	6.2	16.1



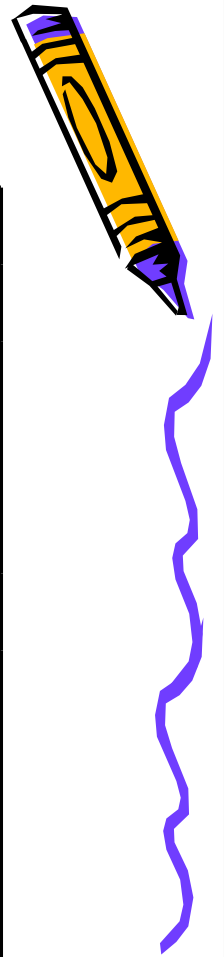
Results for the intermediate model

		Sample 10		Sample 20	
Θ	Method	F10	F20	F10	F20
0	1	4.1	14.2	4.3	14.9
	2	2.7	12.1	3.2	13.2
	3	0.0	0.9	0.1	0.6
	4	2.8	12.4	3.4	13.6
0.4	1	1.1	2.2	1.2	2.3
	2	1.9	13.1	0.8	12.1
	3	0.0	0.2	0.0	0.3
	4	2.1	13.3	0.8	12.2



Results for the completely mixed model

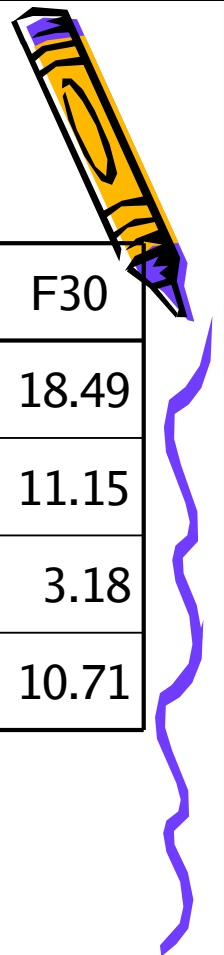
		Sample 10		Sample 20	
Θ	Method	F10	F20	F10	F20
0	1	1.1	6.4	1.5	7.6
	2	0.7	5.6	1.3	7.1
	3	0.0	0.2	0.0	0.1
	4	0.7	5.7	1.3	7.1
0.4	1	1.7	4.7	1.7	4.7
	2	1.8	7.5	1.8	7.4
	3	0.0	0.1	0.0	0.1
	4	1.8	7.6	1.8	7.4



Tests on real data

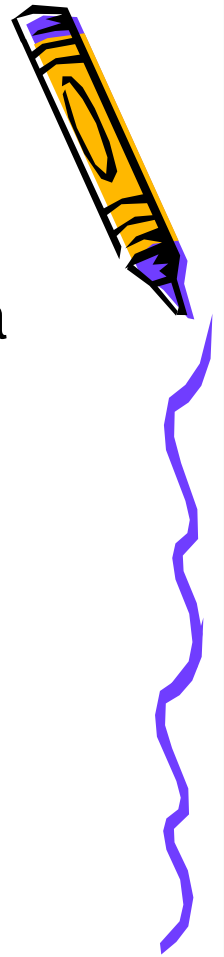
- <http://origo.hu>
- Selected:
 - 1000 users
 - 8321 items
- avg. sample size 50
- without iteration $\Theta = 0$

Method	F10	F20	F30
1	8.13	13.77	18.49
2	3.71	8.13	11.15
3	0.42	1.35	3.18
4	3.50	7.73	10.71



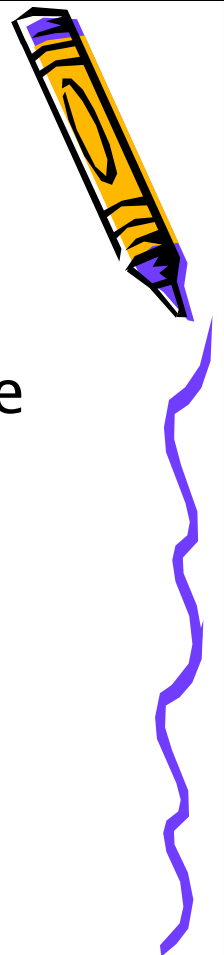
More Applications: Query Recommendations

- Baeza-Yates, Hurtado and Mendoza (2004) suggest a method to make query recommendations based on clustering previous queries.
- Explicitly treat it as a bi-clustering problem.



Applications: Physical and Logical Sessions

- Segment a physical session into distinct logical sessions based on the user behaviour.
- Exploit bi-clustering structure to uncover the hidden logical sessions.



Wisdom of Crowds for Next Generation Exploratory Search

- Paradigm change from query driven search to user-centric information delivery
- Leveraging wisdom of crowds is key
- Biclustering/Multi-clustering is a promising approach.



We've come a Long Way

- "Men go mad in herds, while they only recover their senses slowly one by one."
[Charles McKay, *Madness of Crowds*, 1841.]
- "Madness is the exception in individuals, but the rule in groups" [Nietzsche]
- "I do not believe in the collective wisdom of individual ignorance" [Carlyle]

