Evidence from Behavior

LBSC 796/INFM 719R

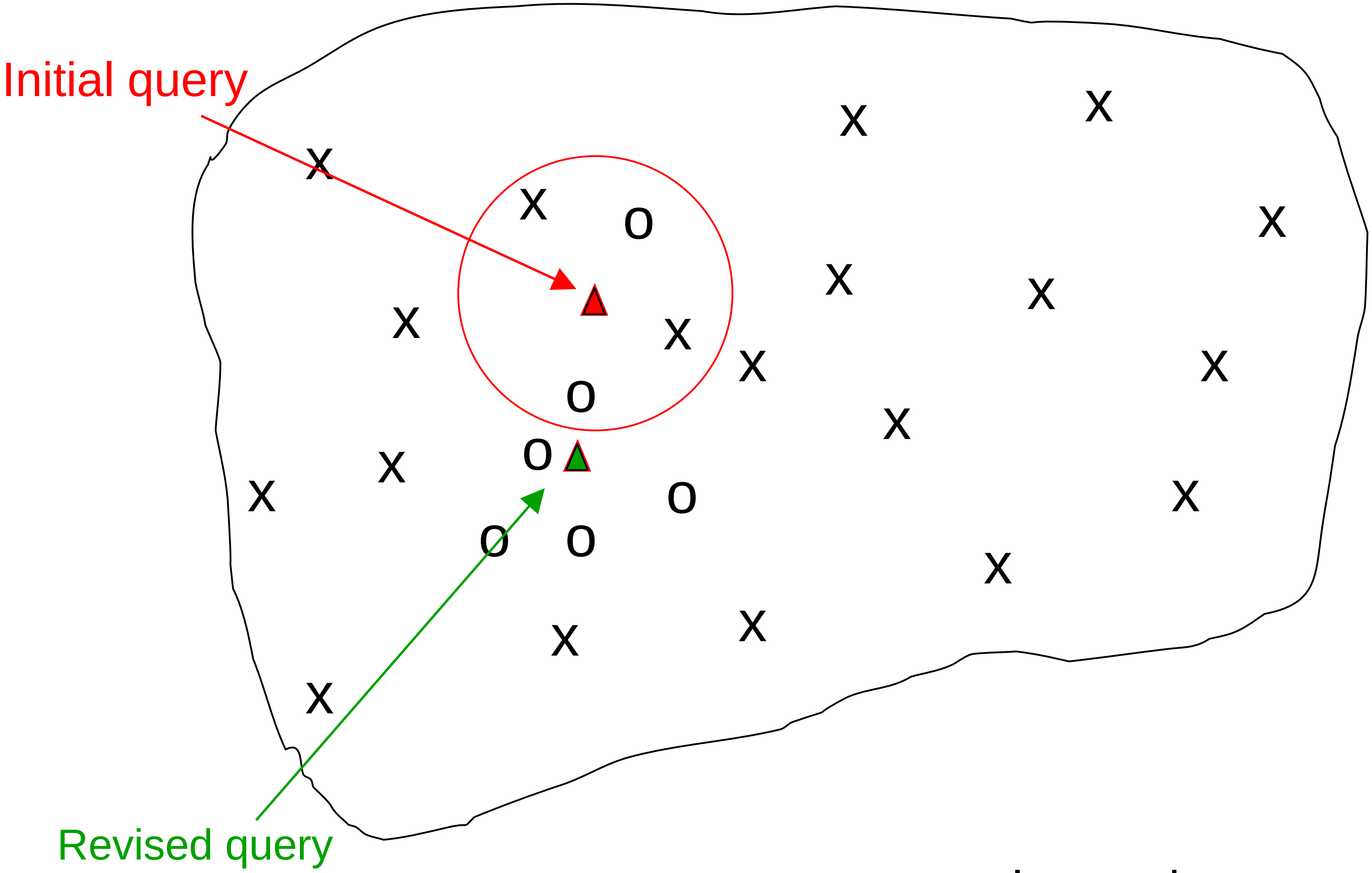
Douglas W. Oard

Session 7, March 16, 2011

Agenda

- Relevance feedback
  - Blind relevance feedback
- “Collaborative” recommendation
- Implicit Feedback
- Query log analysis

Picture of Relevance Feedback



x non-relevant documents

o relevant documents

Rocchio Formula

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

q_m = modified query vector;
q_0 = original query vector;
α, β, γ: weights (hand-chosen or set empirically);
D_r = set of known relevant doc vectors;
D_{nr} = set of known irrelevant doc vectors

Rocchio Example

$$\text{query vector} = \alpha \cdot \text{original query vector}$$
$$+ \beta \cdot \text{positive feedback vector}$$
$$- \gamma \cdot \text{negative feedback vector}$$

Typically, $\gamma < \beta$

Original query	0	4	0	8	0	0	$\alpha = 1.0$	0	4	0	8	0	0	
Positive Feedback	2	4	8	0	0	2	$\beta = 0.5$	1	2	4	0	0	1	(+)
Negative feedback	8	0	4	4	0	16	$\gamma = 0.25$	2	0	1	1	0	4	(-)
							New query	-1	6	3	7	0	-3	

Motivations to Provide Ratings

- Self-interest
  - Use the ratings to improve system's user model
- Economic benefit
  - If a market for ratings is created
- Altruism

“Blind” Relevance Feedback

- Perform an initial search
- Identify new terms strongly associated with top results
  - Chi-squared
  - IDF
- Expand (and possibly reweight) the query

Rating-Based Recommendation

- Use ratings as to describe objects
  - Personal recommendations, peer review, …
- Beyond topicality:
  - Accuracy, coherence, depth, novelty, style, …
- Has been applied to many modalities
  - Books, Usenet news, movies, music, jokes, beer, …

Using Positive Information

	Small World	Space Mtn	Mad Tea Pty	Dumbo	Speedway	Cntry Bear
Joe	D	A	B	D	?	?
Ellen	A	F	D		F	
Mickey	A	A	A	A	A	A
Goofy	D	A		C		
John	A	C	A	C		A
Ben	F	A				F
Nathan	D		A		A	

Source: Jon Herlocker, SIGIR 1999

Using Negative Information

	Small World	Space Mtn	Mad Tea Pty	Dumbo	Speedway	Cntry Bear
Joe	D	A	B	D	?	?
Ellen	A	F	D		F	
Mickey	A	A	A	A	A	A
Goofy	D	A		C		
John	A	C	A	C		A
Ben	F	A				F
Nathan	D		A		A	

Source: Jon Herlocker, SIGIR 1999

Hybrid Systems

- Start with a query
  - Avoids the “cold start” problem
- Obtain some feedback
  - Possibly using “active learning”
- Use the feedback to find other context
  - User-item
  - Item-item

Explicit Feedback: Assumptions

- A1: User has sufficient knowledge for a reasonable initial query

- A2: Selected examples are representative

- A3: The user will give feedback

A1: Good Initial Query?

- Two problems:
  - User may not have sufficient initial knowledge
  - Few or no relevant documents may be retrieved

- Examples:
  - Misspellings (Brittany Speers)
  - Cross-language information retrieval
  - Vocabulary mismatch (e.g., cosmonaut/astronaut)

A2: Representative Examples?

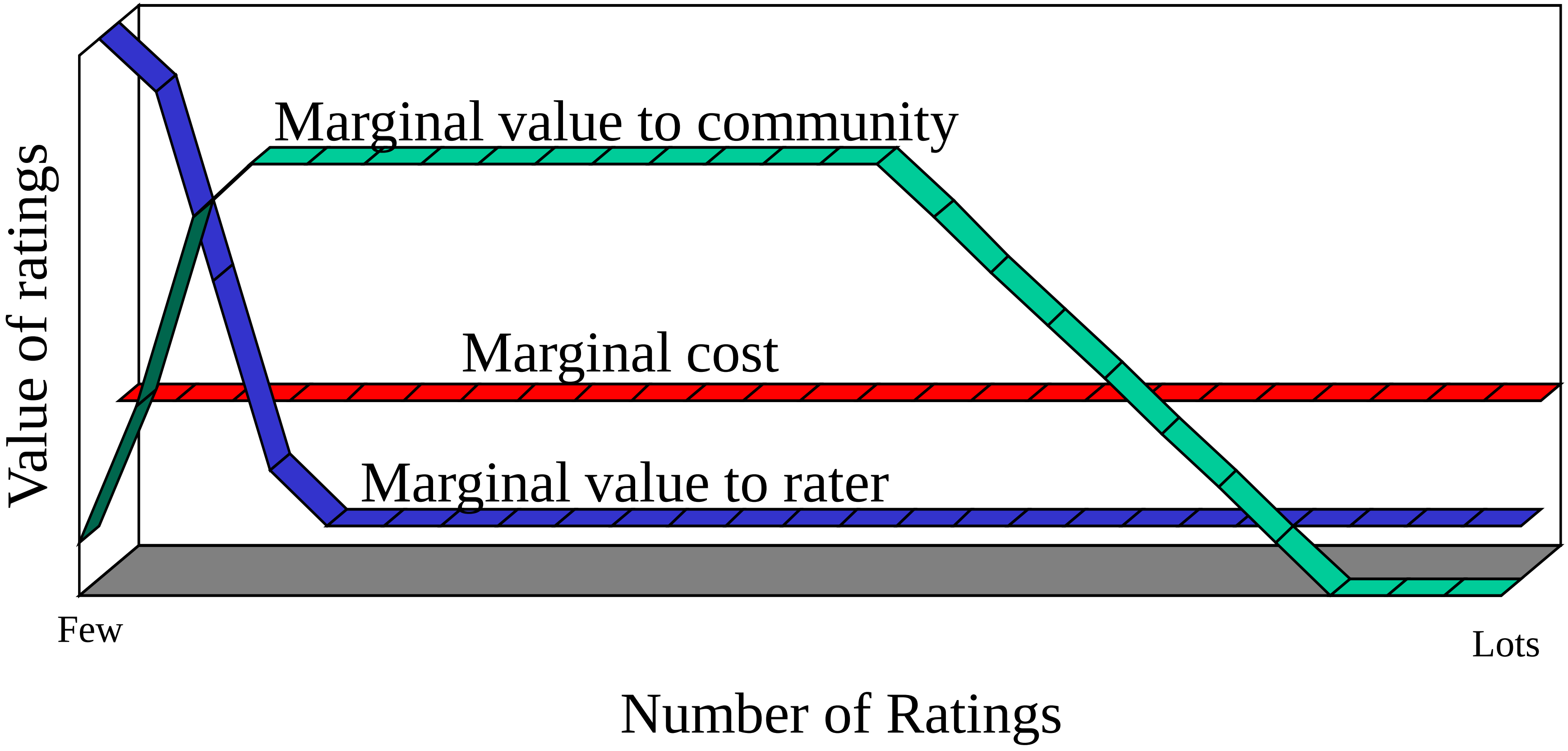
- There may be several clusters of relevant documents

- Examples:
  - Burma/Myanmar
  - Contradictory government policies
  - Opinions

A3: Will People Use It?

- Efficiency
  - Longer queries require more processing time
- Understandability
  - Harder to see why subsequent documents retrieved
- Risk
  - Users are reluctant to provide negative feedback

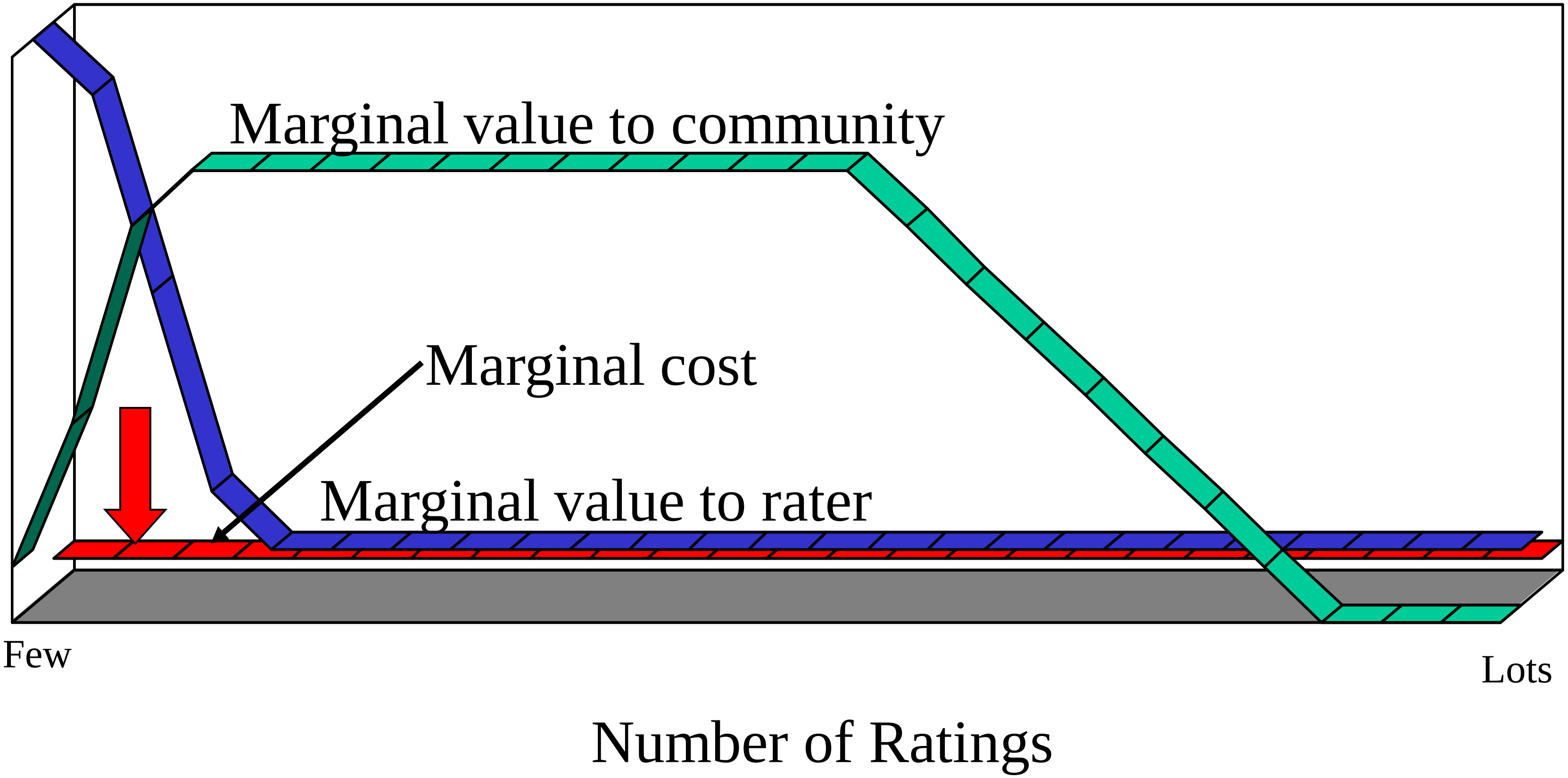
Self-Interest Decreases Over Time



Solving the Cost vs. Value Problem

- Maximize the value
  - Provide for continuous user model adaptation
- Minimize the costs
  - Use implicit feedback rather than explicit ratings
  - Minimize privacy concerns through encryption
  - Build an efficient scalable architecture
  - Limit the scope to noncompetitive activities

Solution: Reduce the Marginal Cost



View	Select	
Listen		
Print	Bookmark	
	Save	
	Purchase	Subscribe
	Delete	
Copy / paste	Forward	
Quote	Reply	
	Link	
	Cite	
Mark up	Tag	Organize
	Publish	
Type		
Edit		

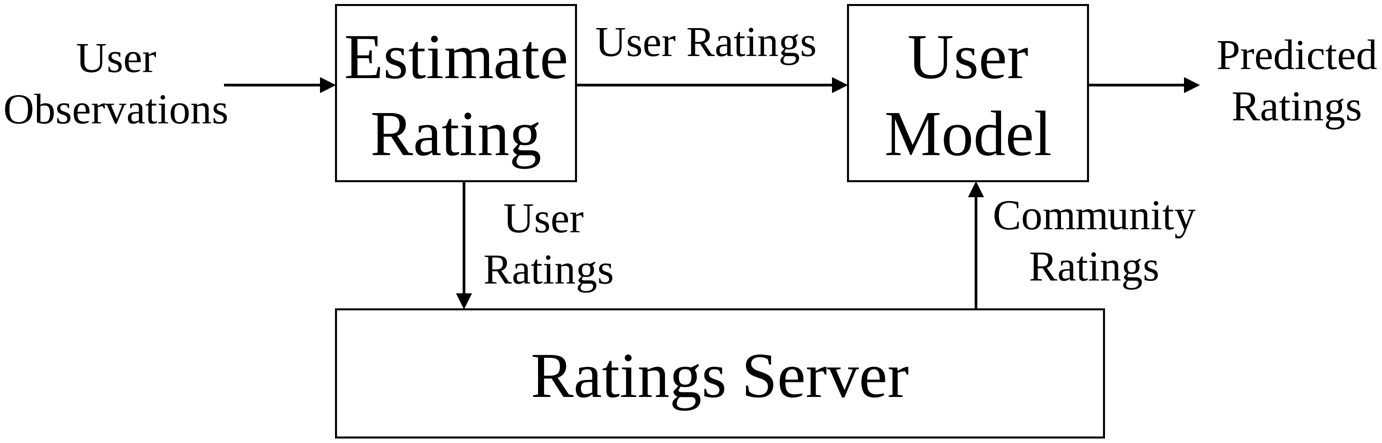
Behavior Category			
Examine	View	Select	
	Listen		
Retain	Print	Bookmark	
		Save	
		Purchase	Subscribe
		Delete	
Reference	Copy / paste	Forward	
	Quote	Reply	
		Link	
		Cite	
Annotate	Mark up	Tag	Organize
		Publish	
Create	Type		
	Edit		

Minimum Scope

Behavior Category

	Segment	Object	Class
Examine	View Listen	Select	
Retain	Print	Bookmark Save Purchase Delete	Subscribe
Reference	Copy / paste Quote	Forward Reply Link Cite	
Annotate	Mark up	Tag Publish	Organize
Create	Type Edit		

Recommending w/Implicit Feedback



User Observations

Estimate Rating

User Ratings

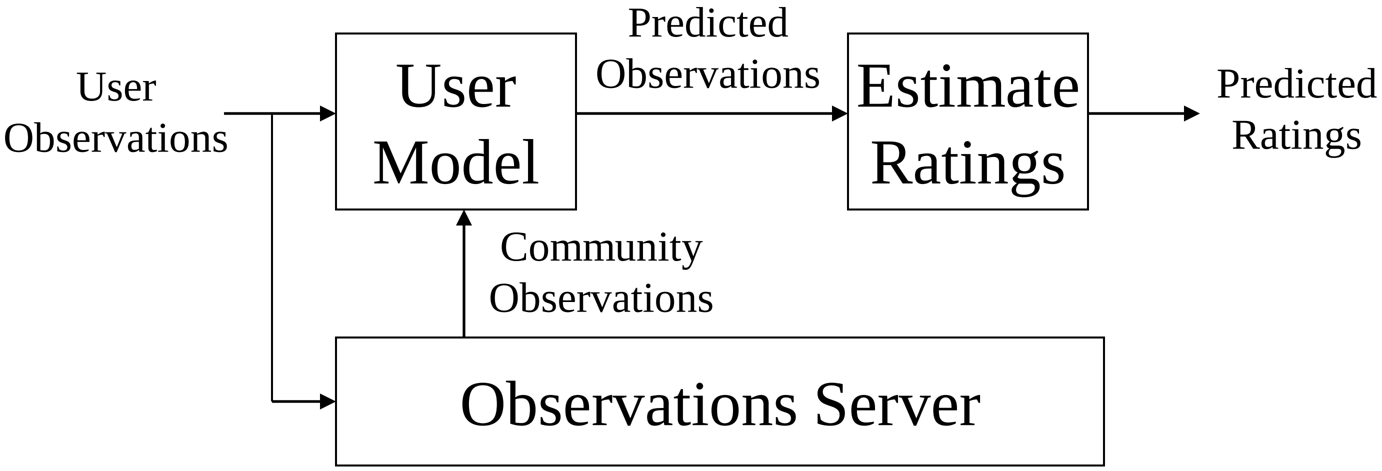
User Model

Predicted Ratings

User Ratings

Community Ratings

Ratings Server



User Observations

User Model

Predicted Observations

Estimate Ratings

Predicted Ratings

Community Observations

Observations Server

Critical Issues

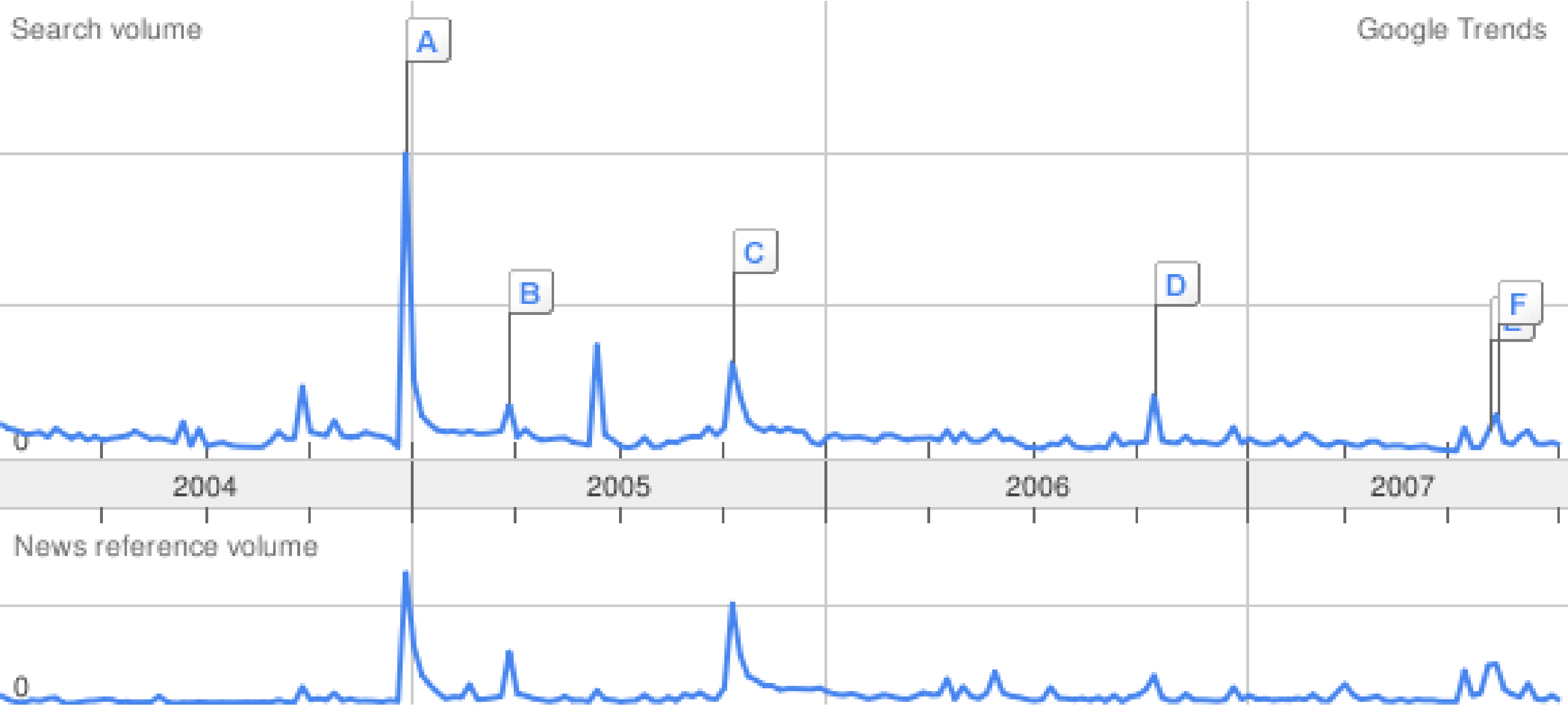
- Protecting privacy
  - What absolute assurances can we provide?
  - How can we make remaining risks understood?
- Scalable rating servers
  - Is a fully distributed architecture practical?
- Non-cooperative users
  - How can the effect of spamming be limited?

Gaining Access to Observations

- Observe public behavior
  - Hypertext linking, publication, citing, …
- Policy protection
  - EU: Privacy laws
  - US: Privacy policies + FTC enforcement
- Statistical assurance of privacy
  - Distributed architecture
  - Model and mitigate privacy risks

Search Engine Query Logs

A: Southeast Asia (Dec 27, 2004)
B: Indonesia (Mar 29, 2005)
C; Pakistan (Oct 10, 2005)
D; Hawaii (Oct 16, 2006)
E: Indonesia (Aug 8, 2007)
F: Peru (Aug 16, 2007)



(3.2)

In this session, the user formulates a series of queries in pursuit of multiple tasks.

In general, the average series of query formulations within a user session can be summarized as a probability matrix (3.4) between the following formulation states:

○	New query
⊕	Add word(s) to query
⊖	Remove word(s) from query
ⓒ	Change word(s) in query
⧁	More results for same query
⧀	Return to a previous query
●	End of session

Timeline (hh:mm:ss)		Query
00:00	○	dail news
01:06	ⓒ	daily news
10:42	○	frito lay
13:48	○	smoking celebrities
14:36	⧁	smoking celebrities
22:18	○	cd reviews
32:48	⧁	cd reviews
40:06	○	bestbuy.com
41:18	○	tower records
47:00	○	money making opportunities
51:42	○	gumball machines
51:54	⧁	gumball machines
57:54	⧁	gumball machines
01:03:48	○	vending opportunities
01:05:48	○	inventions
01:09:00	⧁	inventions
01:18:36	○	patents
01:23:12	⧀	smoking celebrities
01:33:18	○	images.mp3.com
01:33:36	○	www.ajolie.com
01:36:24	○	the sopranos
01:38:30	⧁	the sopranos

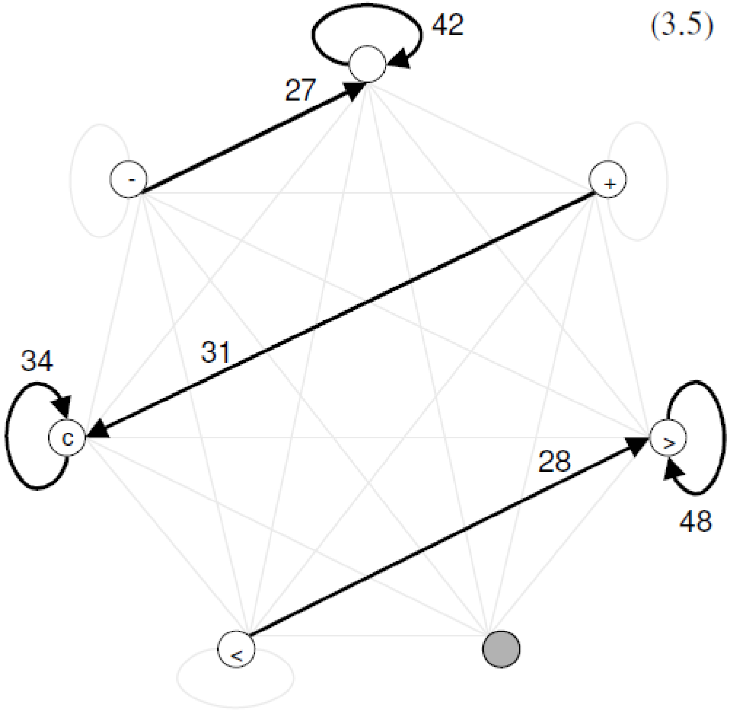
(3.4)

To State

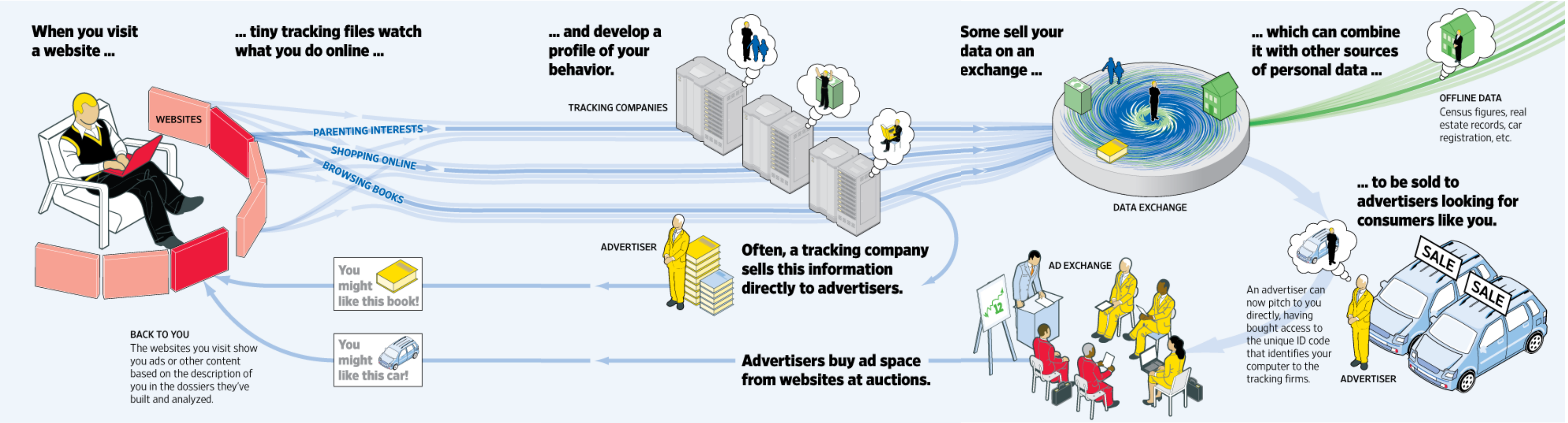
Probability from State

%	○	⊕	⊖	ⓒ	⧁	⧀	●
○	**42**	6	2	15	24	6	5
⊕	25	4	3	**31**	26	8	4
⊖	**27**	18	2	15	26	8	4
ⓒ	20	4	3	**34**	28	6	5
⧁	20	5	1	17	**48**	5	4
⧀	27	4	1	13	**28**	21	6

(3.5)



The Tracking Ecosystem



When you visit a website …

… tiny tracking files watch what you do online …

… and develop a profile of your behavior.

Some sell your data on an exchange …

… which can combine it with other sources of personal data …

… to be sold to advertisers looking for consumers like you.

Often, a tracking company sells this information directly to advertisers.

Advertisers buy ad space from websites at auctions.

BACK TO YOU
The websites you visit show you ads or other content based on the description of you in the dossiers they've built and analyzed.

OFFLINE DATA
Census figures, real estate records, car registration, etc.

An advertiser can now pitch to you directly, having bought access to the unique ID code that identifies your computer to the tracking firms.

http://wsj.com/wtk