An Introduction to Web Science

RuSSIR, Aug 6-10, 2012

Please interrupt at any point!!

Ingmar Weber

ingmar@yahoo-inc.com

Yahoo! Research Barcelona

Course Outline

- Day 1: Introduction to the Introduction
  – Examples, data sets, presentation of the competition
- Day 2: Web Search and Society
  – Demographics, economy and more
- Day 3: Blogs and Twitter
  – Gender, moods, politics, stock market and more
- Day 4: Social Networks and Online Dating
  – Attractiveness, FB&GPA, FB&Personality and more
- Day 5: E-commerce and Marketing Studies
  – Brand congruence, Groupon Effect, social ads

Camera brand congruence in the Flickr social graph

Adish Singla and Ingmar Weber

WSDM 2009

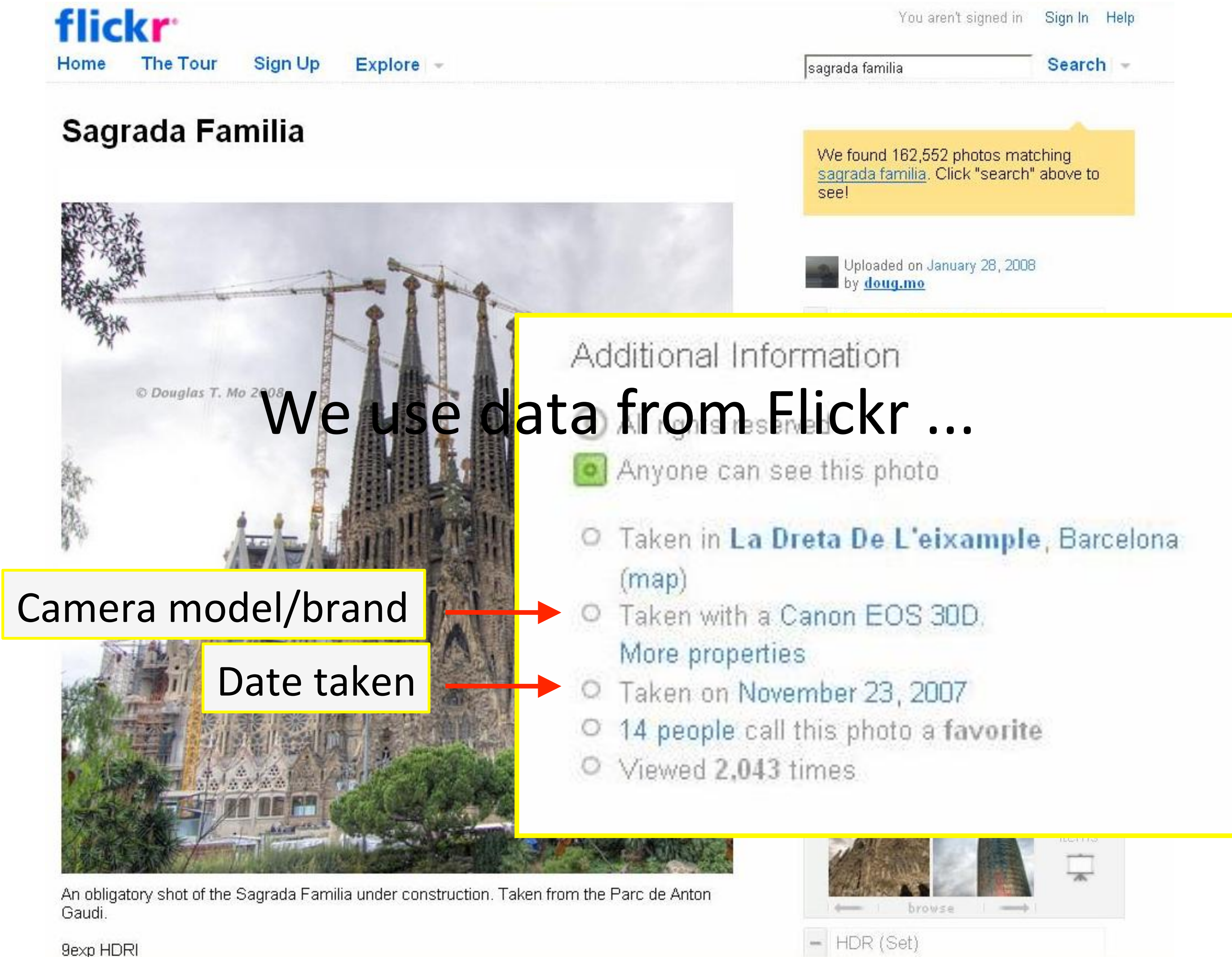
The main research question addressed:

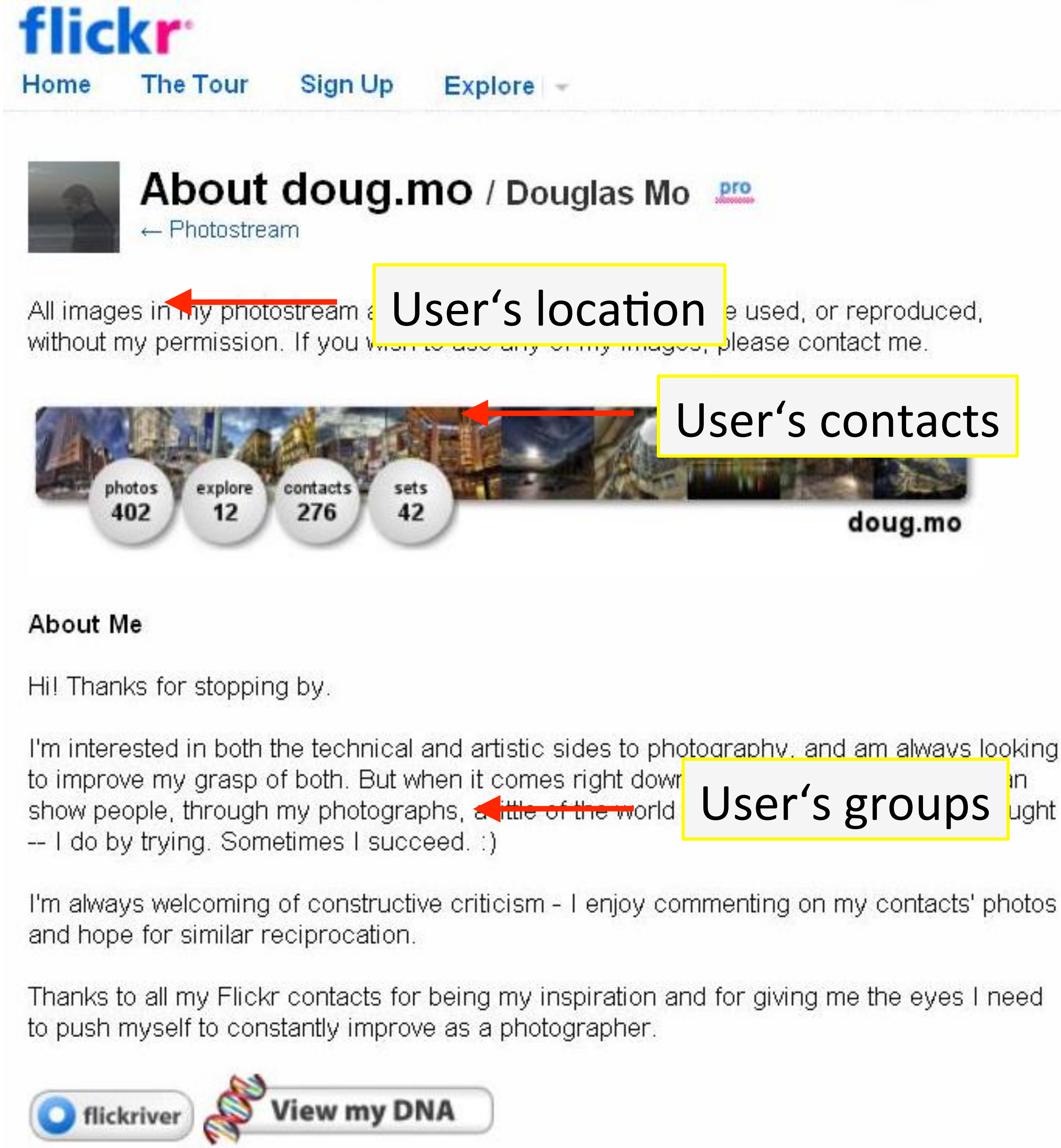
If I use a Sony camera, are my friends more likely to use a Sony camera as well?

Relevant for advertising in social networks.

- whether we are in the same country?
- whether we are close friends?
- whether I use a cheap/expensive camera?
- ....

We use data from Flickr ...




flickr
Home
The Tour
Sign Up
Explore
About doug.mo / Douglas Mo
pro
← Photostream
User‘s location
All images in my photostream a
e used, or reproduced,
without my permission. If you
please contact me.
User‘s contacts
photos 402
explore 12
contacts 276
sets 42
doug.mo
About Me
Hi! Thanks for stopping by.
I'm interested in both the technical and artistic sides to photography, and am always looking
to improve my grasp of both. But when it comes right down
show people, through my photographs, a little of the world
-- I do by trying. Sometimes I succeed. :)
User‘s groups
I'm always welcoming of constructive criticism - I enjoy commenting on my contacts' photos
and hope for similar reciprocation.
Thanks to all my Flickr contacts for being my inspiration and for giving me the eyes I need
to push myself to constantly improve as a photographer.
flickriver
View my DNA

Extracted Information

- Per-image
  - Camera brand
  - Camera model
  - Date taken

- Per-user
  - Location
  - List of contacts
  - List of groups

Data Pre-Processing

- Map camera brand to ID
  - E.g. Minolta = Konika = Konica
- Map camera model to ID
  - E.g. Maxxum 7D = Dynax 7D
- Map location to country ID
  - E.g. California = Canada's neighbor = USA
- Get unique camera brand for users and "buckets"
  - March-May 2006, March-May 2007, March-May 2008
  - Majority voting of (up to) 10 images in a bucket

Data Statistics

- A complete connected component
  - 3.9M users, 67M edges (in summer 2008)
- 1.2M users with brand information
  - 37% Canon, 17% Nikon, 11% Sony, …
- 519k users with country information
  - 39% USA, 9% UK, 5% Canada, …, 27% unmatched
- 11M *directed* edges with brand information
- 1785 models, 96 brands, 168 countries

Methodology: Pairwise Brand Congruence

- Look at user pairs
  – X is in the list of contacts of Y ("friends")
  – X and Y are random users ("baseline")
  – X and Y are friends/random pairs with property Z

- Percentage of *congruent* pairs
  – Congruent = same brand used
  – High congruence itself is **not** enough
  – Is the percentage for friends higher than for baseline

Dependence on Friendship and Country

	%-age congruent in April 2008	
	Random	**Friends**
Country ignor		

Friendship matters ...
... more than country.

Dependence on Closeness of Friendship

“close” = similar interests = similar groups joined

$X \in \{G_1, G_2, G_4\}$, $0.0 < G_J < 0.24$	Y ...			$\{..., G_1, G_3, G_4, G_5, G_6\}$ 1.0		$G_J = 1.0$, $\| = 2/6$
27%				7%		27%

Groups are irrelevant.

“close” = mutual friends

27%			

Mutuality is irrelevant.

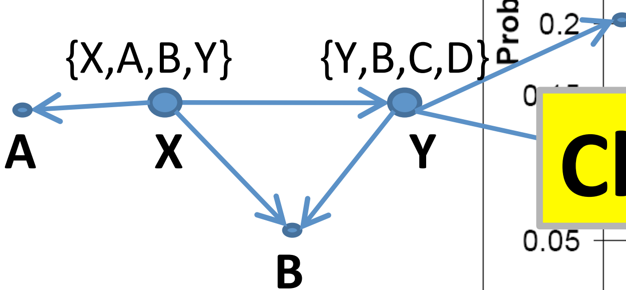
“close” = few friends (up to five)

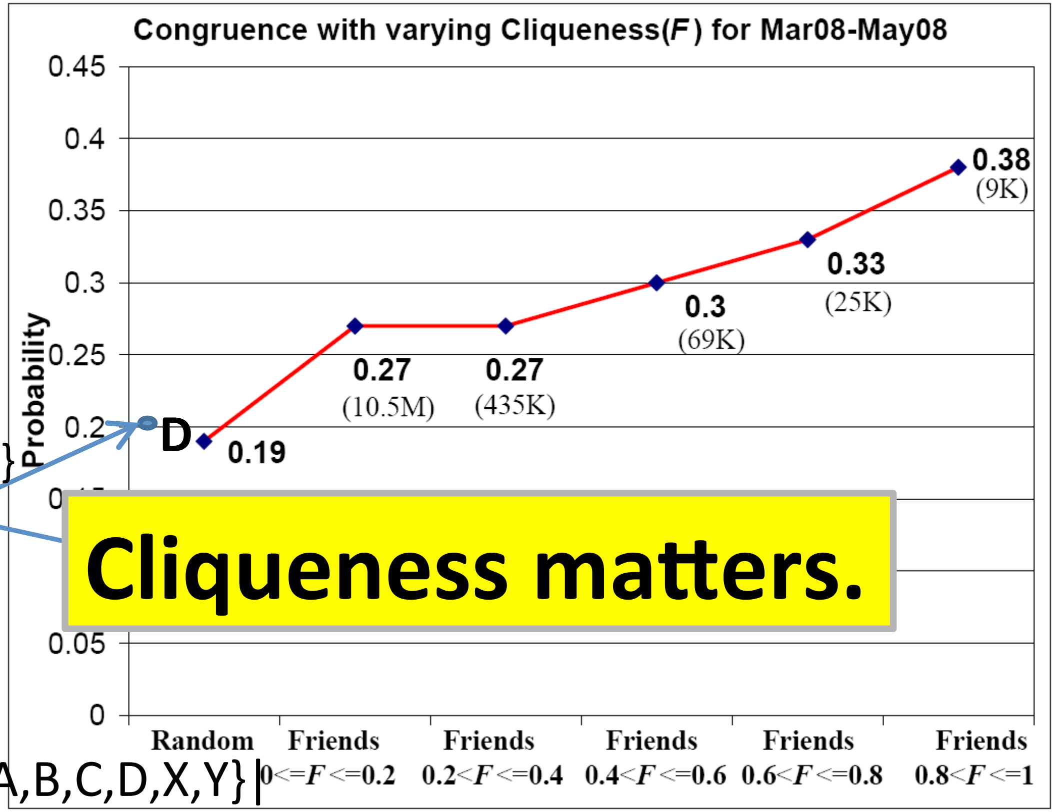
	small-small	small-large	large-small	large-large
country ignore			%)	27% (21%)

Friendship size matters.

Dependence on Closeness of Friendship

"close" = cliqued





Cliqueness matters.

$F_J(X,Y) = |\{B,Y\}| / |\{A,B,C,D,X,Y\}|$
$= 2/6$

	$0 \cdot F_J \cdot 0.5$	$0.5 < F_J \cdot 1.0$
country ignored	27%	33%
Same country	big difference	33%
diff. country	28%	24%

Dependence on Camera Type

Point & Shoot (P&S) = cheap, used by "beginner" users

Digital Single Lens Reflex (DSLR) = expensive, used by "expert" users

no huge difference

no huge difference

		S - DSLR	DSLR –	
country ignored	26% (20%)	20% (19%)	20% (19%)	47% (42%)

Camera type matters.

			&S	DSLR – DSLR
cliquen. ignored	26%	20%	20%	47%

big difference

big difference

“Triggering” of Brand/Model Changes

- Given a user changes her model 2007 -> 2008
  - 54% high / 51% low cliqueness also change
  - 48% of random users change
- Given a model change of user *and* friend
  - rand

There seems to be some “triggering”.

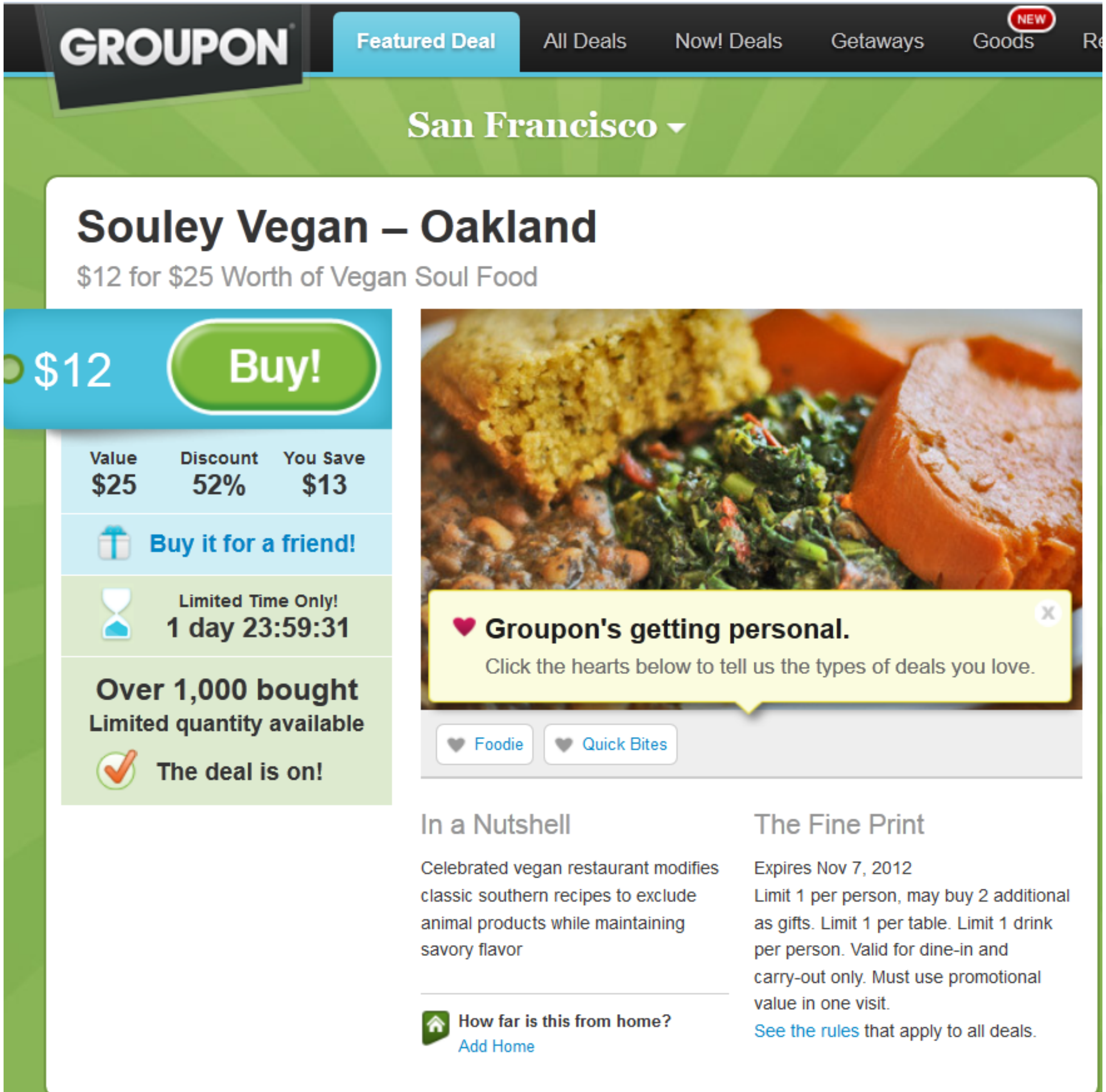
  - c.f. 33% congruent high cliqueness friends in 2008
- Given a model change of random users
  - 20% change to same brand
  - c.f. 19% congruent in 2008
- Country information only added 1-2%

The Groupon Effect on Yelp Ratings: A Root Cause Analysis

John Byers, Michael Mitzenmacher and Georgios Zervas

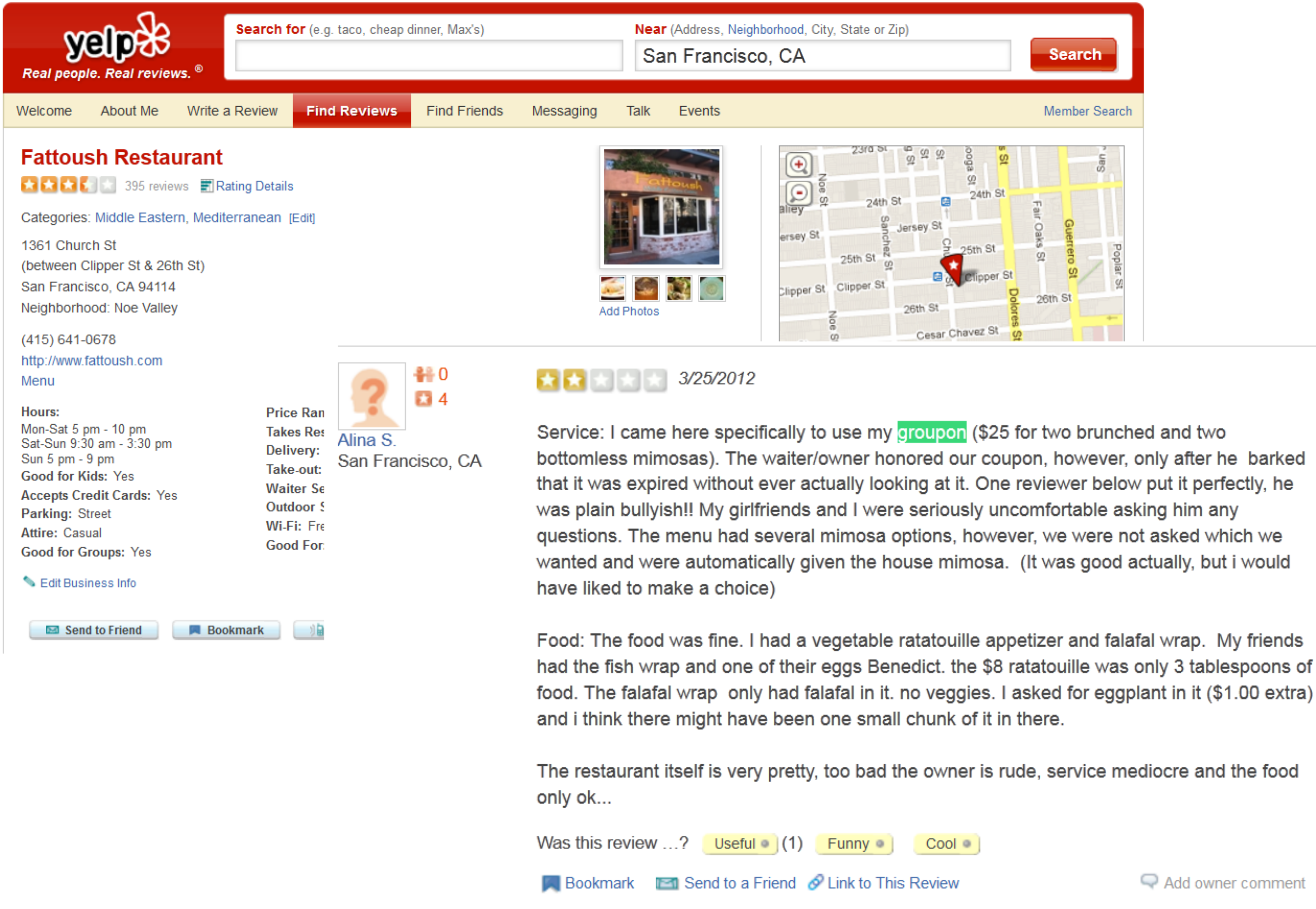
EC'12

Groupon



GROUPON
Featured Deal
All Deals
Now! Deals
Getaways
NEW
Goods
San Francisco
Souley Vegan – Oakland
$12 for $25 Worth of Vegan Soul Food
$12
Buy!
Value $25
Discount 52%
You Save $13
Buy it for a friend!
Limited Time Only!
1 day 23:59:31
Over 1,000 bought
Limited quantity available
The deal is on!
Groupon's getting personal.
Click the hearts below to tell us the types of deals you love.
Foodie
Quick Bites
In a Nutshell
Celebrated vegan restaurant modifies classic southern recipes to exclude animal products while maintaining savory flavor
How far is this from home?
Add Home
The Fine Print
Expires Nov 7, 2012
Limit 1 per person, may buy 2 additional as gifts. Limit 1 per table. Limit 1 drink per person. Valid for dine-in and carry-out only. Must use promotional value in one visit.
See the rules that apply to all deals.

Yelp


yelp
Real people. Real reviews.®
Search for (e.g. taco, cheap dinner, Max's)
Near (Address, Neighborhood, City, State or Zip)
San Francisco, CA
Search
Welcome
About Me
Write a Review
Find Reviews
Find Friends
Messaging
Talk
Events
Member Search
Fattoush Restaurant
395 reviews
Rating Details
Categories: Middle Eastern, Mediterranean [Edit]
1361 Church St
(between Clipper St & 26th St)
San Francisco, CA 94114
Neighborhood: Noe Valley
(415) 641-0678
http://www.fattoush.com
Menu
Hours:
Mon-Sat 5 pm - 10 pm
Sat-Sun 9:30 am - 3:30 pm
Sun 5 pm - 9 pm
Good for Kids: Yes
Accepts Credit Cards: Yes
Parking: Street
Attire: Casual
Good for Groups: Yes
Price Ran
Takes Res
Delivery:
Take-out:
Waiter Se
Outdoor S
Wi-Fi: Fre
Good For:
Edit Business Info
Send to Friend
Bookmark
Add Photos
23rd St
24th St
Jersey St
25th St
Clipper St
26th St
Cesar Chavez St
Sanchez St
Noe St
Church St
Dolores St
Fair Oaks St
Guerrero St
Poplar St
0
4
Alina S.
San Francisco, CA
3/25/2012
Service: I came here specifically to use my groupon ($25 for two brunched and two bottomless mimosas). The waiter/owner honored our coupon, however, only after he barked that it was expired without ever actually looking at it. One reviewer below put it perfectly, he was plain bullyish!! My girlfriends and I were seriously uncomfortable asking him any questions. The menu had several mimosa options, however, we were not asked which we wanted and were automatically given the house mimosa. (It was good actually, but i would have liked to make a choice)
Food: The food was fine. I had a vegetable ratatouille appetizer and falafal wrap. My friends had the fish wrap and one of their eggs Benedict. the $8 ratatouille was only 3 tablespoons of food. The falafal wrap only had falafal in it. no veggies. I asked for eggplant in it ($1.00 extra) and i think there might have been one small chunk of it in there.
The restaurant itself is very pretty, too bad the owner is rude, service mediocre and the food only ok...
Was this review ...? Useful (1) Funny Cool
Bookmark
Send to a Friend
Link to This Review
Add owner comment

Ratings Decline – Why?

- Their prior work
  - "negative side effect for merchants selling Groupons is that, on average, their Yelp ratings decline significantly"
- Why does this happen?
  - Critical users?
  - Users outside their normal "sphere"?
- Their claim
  - "reviews from Groupon users are lower on average because such reviews correspond to real, unbiased customers"

Dataset

- Groupon.com and Yelp.com
  - Groupon: 16.7k deals during Jan-Jul 2011
  - 5,472 Groupon businesses identified with Yelp
  - Get **all** reviews of users reviewing a Groupon Bus.
  - 7.1M reviews for 942k business
  - Split reviews for seed business into two sets
    - Given by users with the term “groupon” in any review
    - By the other users

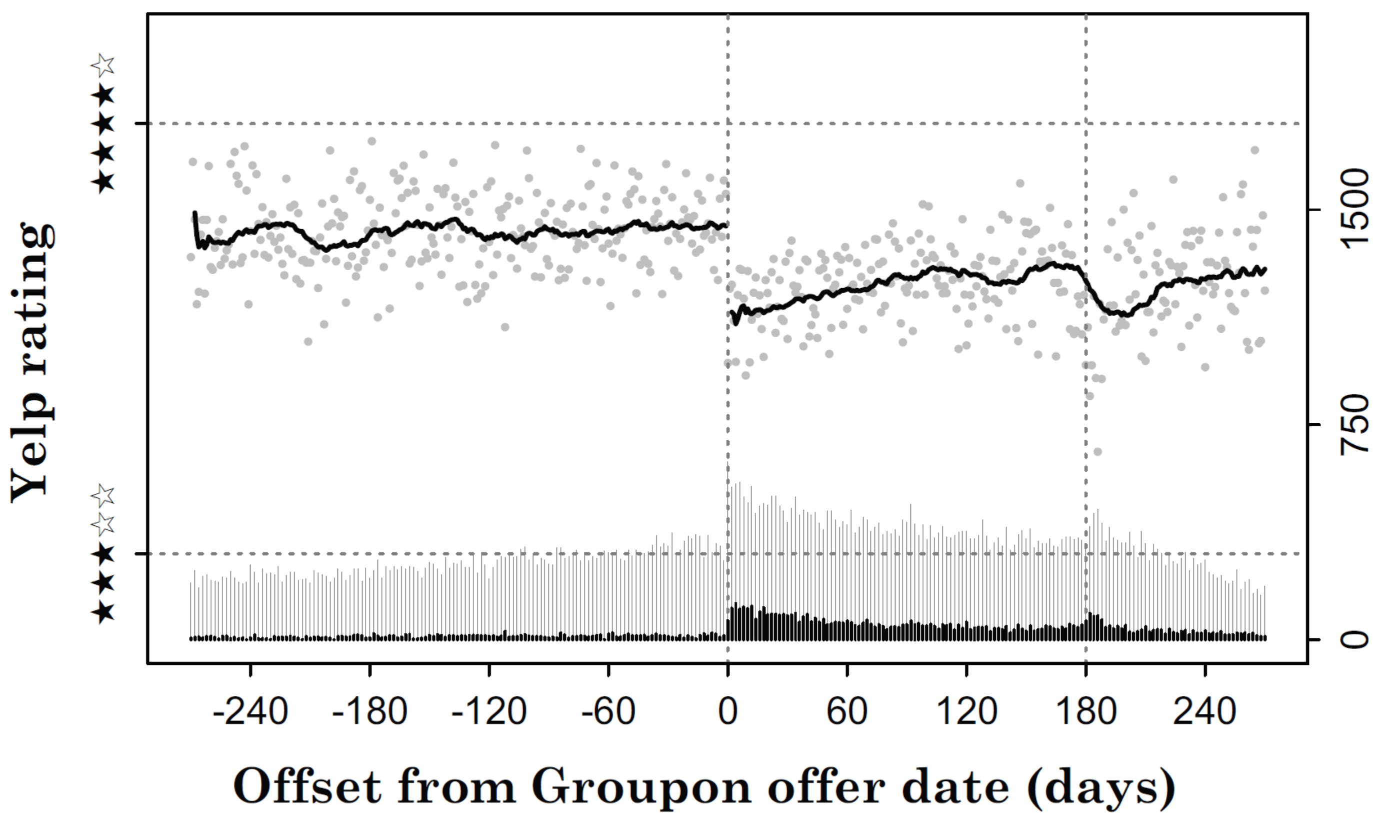
Two Different Kinds of Reviewers

	Yelping Since	Friends	Fans	Reviews	Firsts	Count
Groupon user	2009-06-27	44.94	4.38	89.60	7.19	21,020
	(506.18)	(144.28)	(16.74)	(160.34)	(29.40)	
Not a Groupon user	2009-06-01	24.43	1.92	44.25	3.72	127,946
	(530.01)	(106.62)	(12.49)	(88.57)	(19.32)	

- Groupon users are “Mavens” (= “information specialists”) in “The Tipping Point”-sense, Malcolm Gladwell

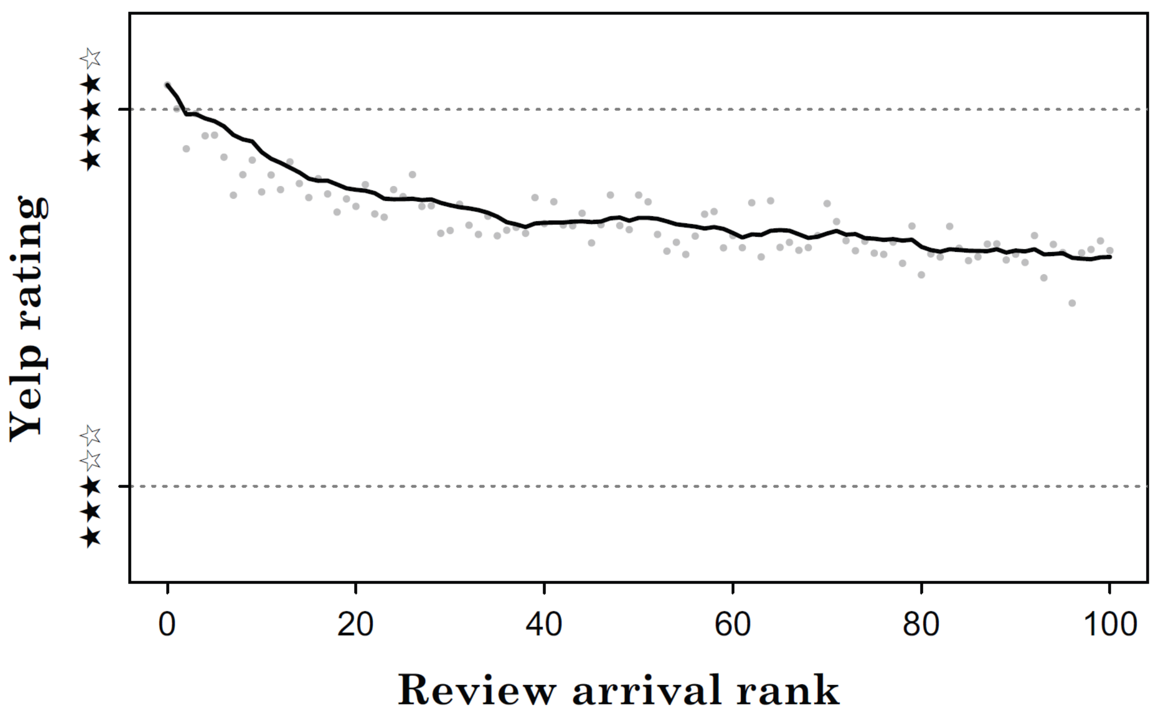
The Groupon Effect

- Groupon reviews: average rating 3.27 stars
- Non-Groupon reviews: av. Rating 3.73 stars



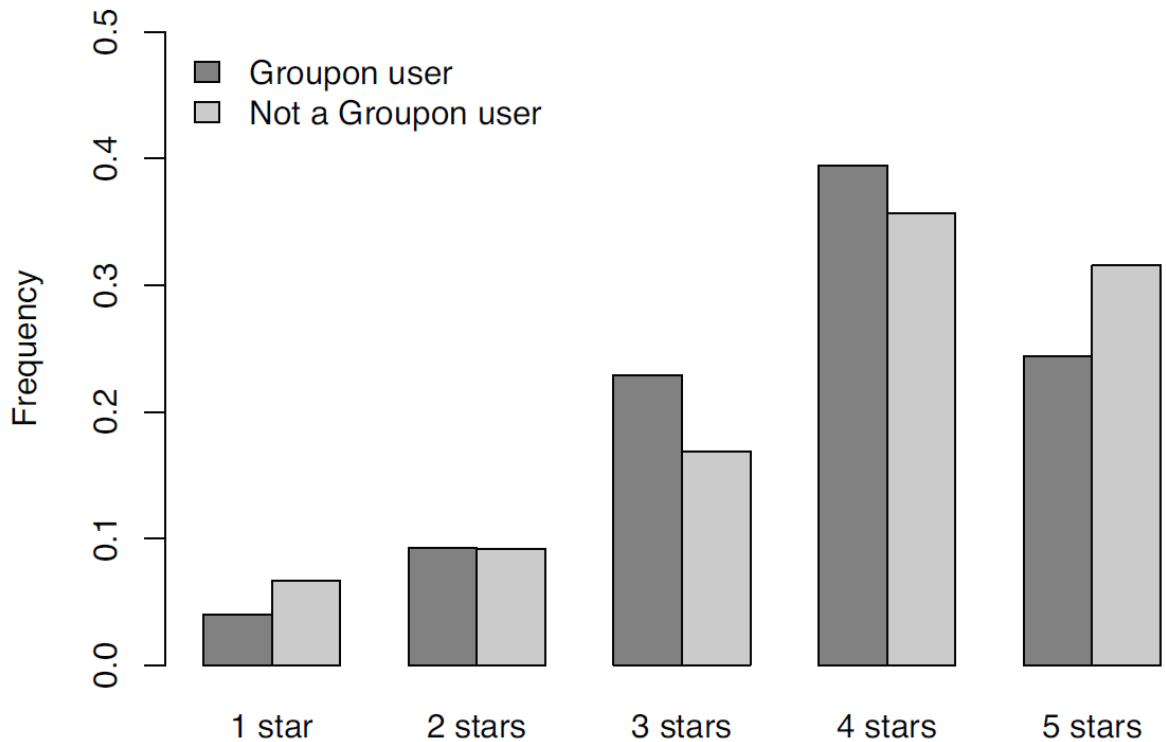
Hypothesis 1: Intrinsic Decline

- It is well known that review scores fall over time, and this is the effect seen (largely independent of Groupon)



Hypothesis 2: Critical Reviewers

- Groupon users are more critical than their peers

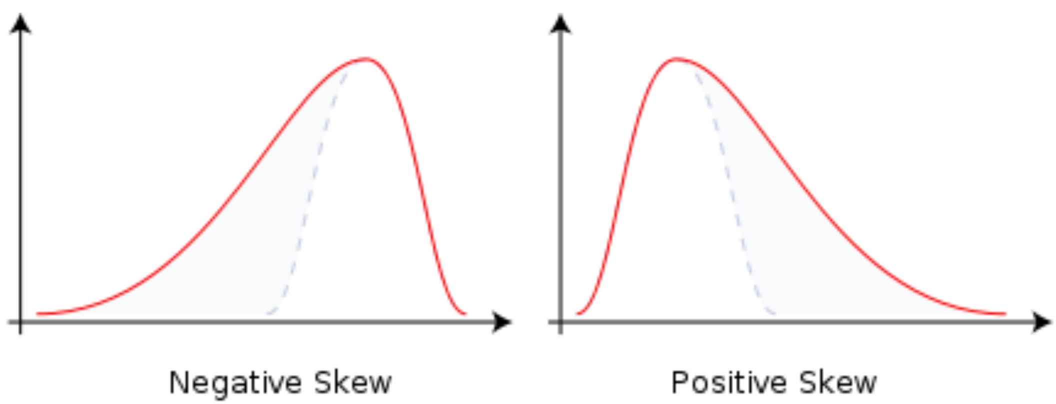


(a) Non-Groupon businesses

Aver. 3.71 by Groupon users (for non-Groupon business)
Aver. 3.76 by non-Groupon users (for non-Groupon business)

Hypothesis 3: Bad Businesses

- Merchants who feel compelled to offer a Groupon are desperate, or in trouble anyway
- FTD flower "bait and switch" scheme
- More skewed? Some really bad guys?



$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

- Observed a slightly more negative skew

Hypothesis 4: Experimentation

- Groupon users are often experimenting when they purchase a Groupon, trying a business category they would not normally frequent
- Look at categorization

Table II: Summary statistics of consumer experimentation.

Groupon user	Groupon mention	Category match?		ZIP match?	
		Yes	No	Yes	No
False	False	70%	30%	68%	32%
True	False	84%	16%	80%	20%
True	True	67%	33%	66%	34%

Hypothesis 5: Artificial Reviews

- Groupon reviews are a more realistic baseline, because the rest of the reviews contain a higher fraction of artificially laudatory reviews

Table III: Percentage of filtered reviews for Groupon vs. non-Groupon users.

Groupon user	Groupon mention	Reviews			
		Visible	Filtered	Filtered pct.	Avg. Rating
False	False	4,837	723	14.95%	3.79
True	False	6,496	707	10.88%	3.58
True	True	175	19	10.86%	3.28

Modeling the Generation of Yelp Rating Scores

- Probit model: $y_{ij}^* = \mathbf{x}_{ij}'\boldsymbol{\beta} + \epsilon_{ij}$

$y_{ij} \in \{1,2,3,4,5\}$ $\quad \epsilon_{ij} \sim \mathcal{N}(0,1)$

$$y_{ij} = \begin{cases} 1 & \text{if } y_{ij}^* \le \kappa_1, \\ 2 & \text{if } \kappa_1 < y_{ij}^* \le \kappa_2, \\ 3 & \text{if } \kappa_2 < y_{ij}^* \le \kappa_3, \\ 4 & \text{if } \kappa_3 < y_{ij}^* \le \kappa_4, \\ 5 & \text{if } \kappa_4 < y_{ij}^*, \end{cases}$$

$$Pr[y_{ij} \le n] = \Phi(\kappa_n - \mathbf{x}'\boldsymbol{\beta})$$

$$Pr[y_{ij} = n] = \Phi(\kappa_n - \mathbf{x}_{ij}'\boldsymbol{\beta}) - \Phi(\kappa_{n-1} - \mathbf{x}_{ij}'\boldsymbol{\beta})$$

- probit = inverse CDF for N(0,1)
- Use maximum-likelihood approach for fitting

Modeling the Generation of Yelp Rating Scores

$$\mathbf{probit}(\mathbf{Pr}[y_{ij} \le n]) = \kappa_n - C_{in} - B_{jn} - R_{ijn}$$

$$C_{in} = \gamma_{1n} \times \mathbf{Groupon\ user}_i,$$

$$B_{jn} = \sum_{p=2}^{\#cities} \beta_{2p} \times \mathbf{Deal\ city}_j + \sum_{q=2}^{\#categ.} \beta_{3q} \times \mathbf{Deal\ category}_j$$

$$+ \gamma_{2n} \times \mathbf{During\ Groupon}_j + \gamma_{3n} \times \mathbf{Post\ Groupon}_j,$$

$$R_{ijn} = \gamma_{4n} \times \mathbf{Groupon\ mention}_{ij} + \gamma_{5n} \times \mathbf{Review\ rank}_{ij}.$$

Marginal effect

$$E_x\left[\frac{\partial \mathbf{Pr}[y_{ij} = n | x_{ij}]}{\partial x_{ij}^{(m)}}\right]$$

Average marginal effects of receiving a specific Yelp rating.

	Yelp rating				
	1	2	3	4	5
Groupon mention	10.223%	3.802%	-2.242%	-7.388%	-4.394%
Groupon user	-4.436%	1.095%	6.544%	6.398%	-9.601%
During Groupon deal	3.577%	0.778%	-2.544%	-4.017%	2.206%
Post Groupon deal	2.791%	0.575%	-1.737%	-3.069%	1.440%
Review rank	-0.006%	0.002%	0.008%	0.010%	-0.013%

So, the Reason is …

- More analysis in the paper
- Punchline:

  "While there remain challenges in trying to exactly quantify the different issues at play, we have shown that a combination of poor business behavior, Groupon user experimentation, and an artificially high baseline all play a role."

Social Influence in Social Advertising: Evidence from Field Experiments

Eytan Bakshy, Dean Eckles, Rong Yan and Itamar Rosenn

EC'12

Correlation or Causation

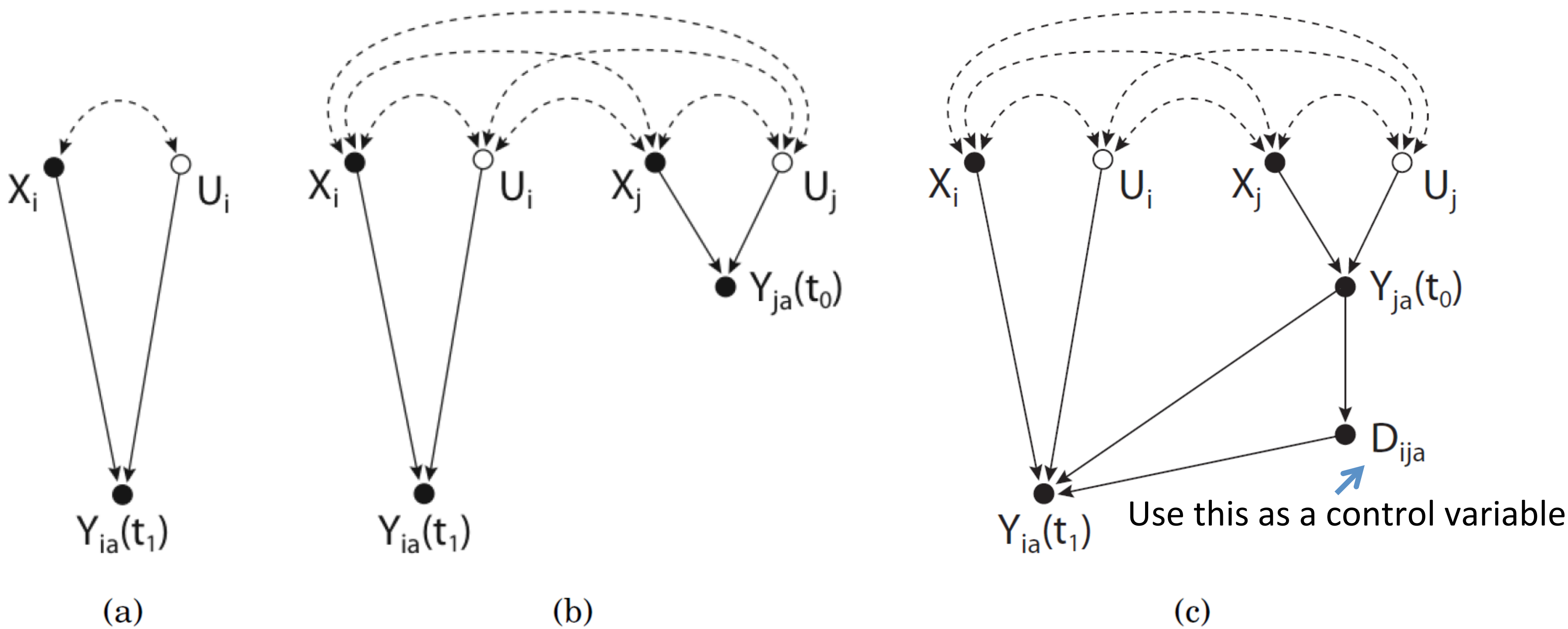


Fig. 1. Causal relationships in consumer responses to advertising. Solid lines indicate cause-and-effect relationships. Dashed lines indicate that variables are correlated in some (possibly unknown) way. (*a*) Responses are caused by observed and unobserved individual characteristics. (*b*) Responses may be correlated with peers' responses even when there is no social influence. (*c*) Responses can be explained both by social influence and correlation among peer characteristics. Here one mechanism for social influence, among other possible mechanisms, is the inclusion of social cues, D_{ija}, in the ad.

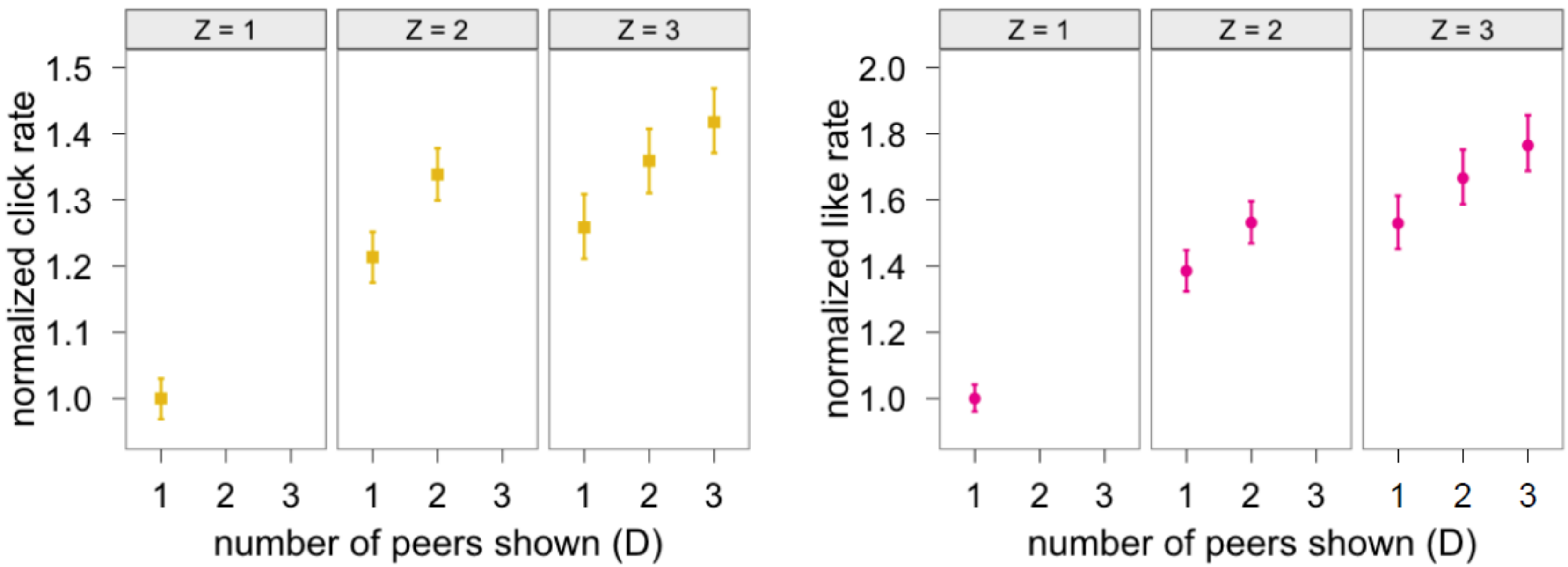
Assessing Response Rates

- Response rates are not i.i.d.
- Observing 100,000 impressions for 10,000 users on 1,000 ads gives optimistic error bounds
- Apply weighted bootstrap sampling on *pairs*
  - Each user and ad is given a Poisson(1) weight
  - Multiplied, sampled and repeated N times
  - Gives conservative estimates of the variance
  - https://github.com/deaneckles/multiway_bootstrap

Influence of Multiple Peers

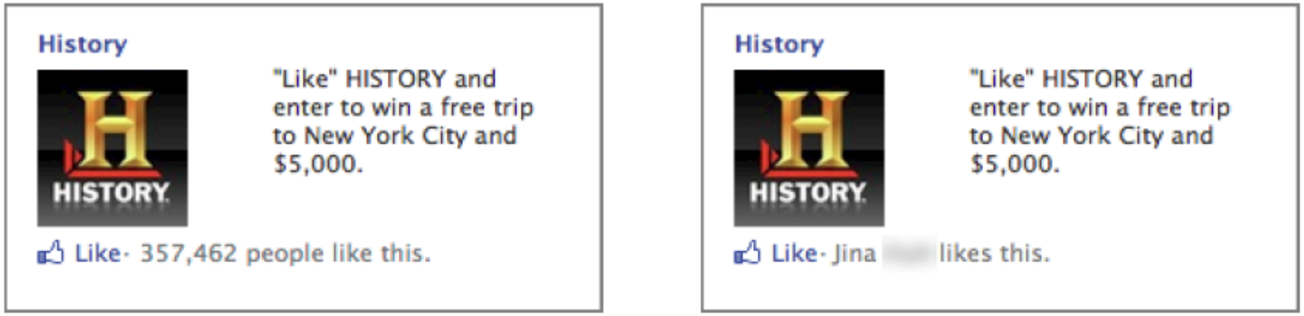
(a) (b) (c)

Fig. 2. Experimental treatment for sponsored story ad units in Experiment 1. Figure illustrates the three possible treatment conditions for users with three peers ($Z_{ia} = 3$) who are affiliated with the sponsored page. (a) $D_{ia} = 1$ (b) $D_{ia} = 2$ (c) $D_{ia} = 3$.



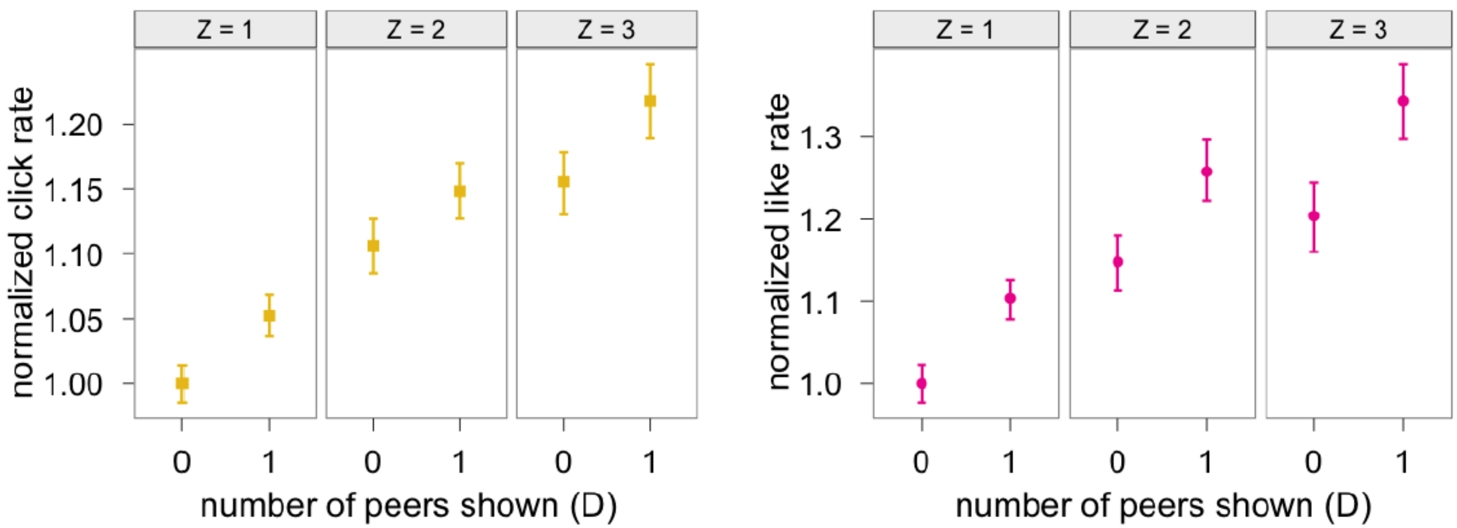
(a) (b)

Influence of Minimal Social Cues



(a) (b)

Fig. 4. The two treatment conditions for social ads in Experiment 2. Subjects who are to be exposed to ads with at least one affiliated peer are randomly assigned to see either (a) general information about the total number of affiliated individuals ($D_{ia} = 0$) or (b) a minimal social cue featuring one affiliated peer ($D_{ia} = 1$).



(a) (b)

Tie Strength

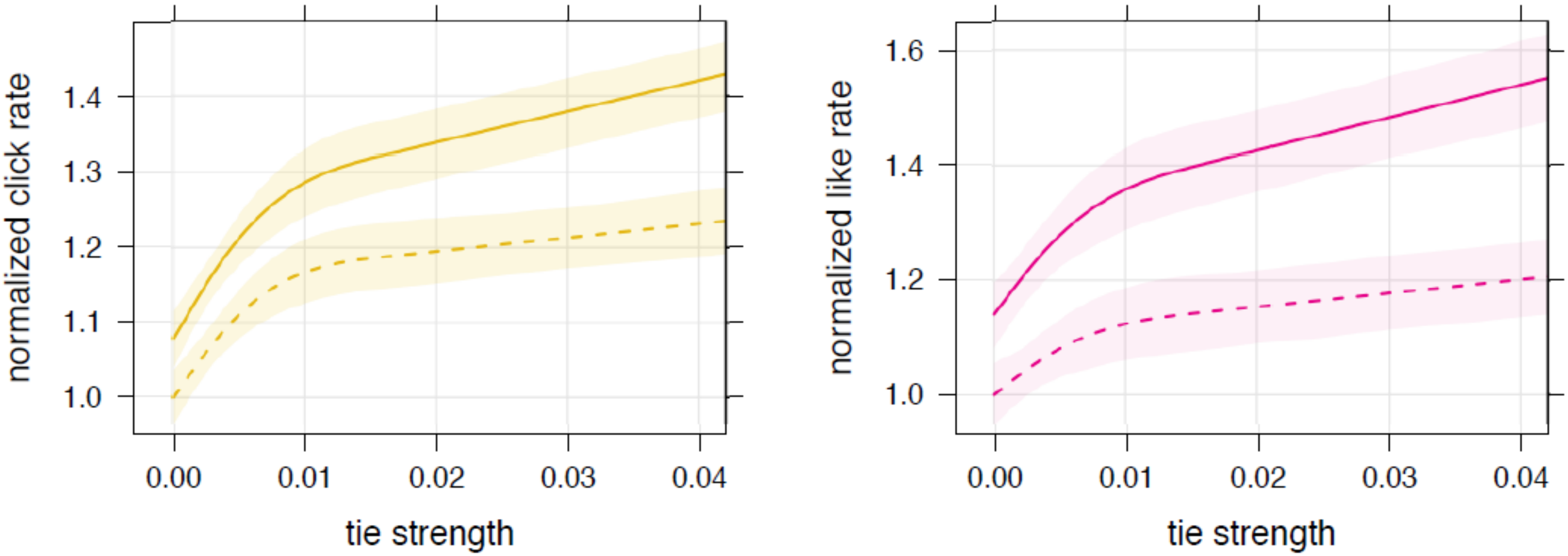
- Directed tie strength $W_{ij} = C_{ij}/C_{i\bullet}$
- Percentile-transformation for user activity $q(C_{i\bullet})$

$$Y_{ija} \sim \alpha + \delta D_{ia} + \tau f(W_{ij}) + \eta D_{ia} \cdot f(W_{ij}) + \gamma q(C_{i\bullet}) + \lambda q(C_{i\bullet}) \cdot f(W_{ij})$$

where f is a natural spline basis expansion for measured tie strength with knots at the second and third quartiles of measured tie strength over all impressions.

- Spline: an approximation to a noisy, discrete curve
- "natural": smoothest curve with exact fit
- "smooth": small absolute second derivative

Influence of Tie Strength



(a) (b)

Fig. 7. Estimated average response as a function of tie strength between the user and the single affiliated peer. Action rates increase with tie strength both in the presence ($D = 1$, solid) and absence ($D = 0$, dashed) of the minimal social cue featuring the affiliated peer. Each plot shows model fits (via Equation 1) for users at the median total communication count (i.e., $q(C_{i\bullet}) = 0.5$), ranging from zero to the 90th percentile of tie strength. Shaded regions are 95% bootstrapped confidence intervals of the predicted response rate, which are generated by fitting the model to $R = 500$ bootstrap replicates of the data.

Reminder:

Competition

Research Proposal Presentations

- Proposal 1: XXX
- Proposal 2: XXX
- Proposal 3: XXX
- Two minutes maximum!

Applausometer Results

- Proposal 1: XXX
- Proposal 2: XXX
- Proposal 3: XXX

Questions?

End of Day 5

ingmar@yahoo-inc.com