How can you tell if a discussion is interesting?

Characterizing and curating conversation threads

Lars Backstrom
Facebook
lars@fb.com

Jon Kleinberg
Cornell
kleinber@cs.cornell.edu

Lillian Lee
Cornell
llee@cs.cornell.edu

Cristian Danescu-Niculescu-Mizil
Stanford/Max Planck Institute SWS
cristiand@cs.stanford.edu

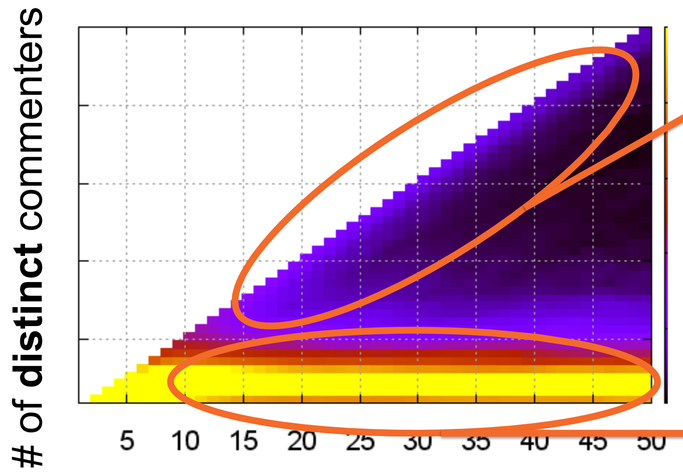
TWO PREDICTION PROBLEMS

(1) LENGTH, aka volume (2) RE-ENTRY

Images: www.facebook.com; www.wikipedia.org; Chase Mitchell, College Humor, 2010

This heatmap reveals a key bimodality:

As thread length increases, there's… **EXPANSIONARY threads** vs **FOCUSED threads**



most commenters appear just once

few commenters, who re-enter a lot

Commenter may want to be kept informed about this thread.

Threads created with fakebookquotes.appspot.com

Aside: It inspires a new generative model.

A standard Chinese-restaurant-style model *can't* generate this bimodality.

We show by simulation that it *can* be generated by a **model inspired by the theory of nonlinear urn processes.**

The details:

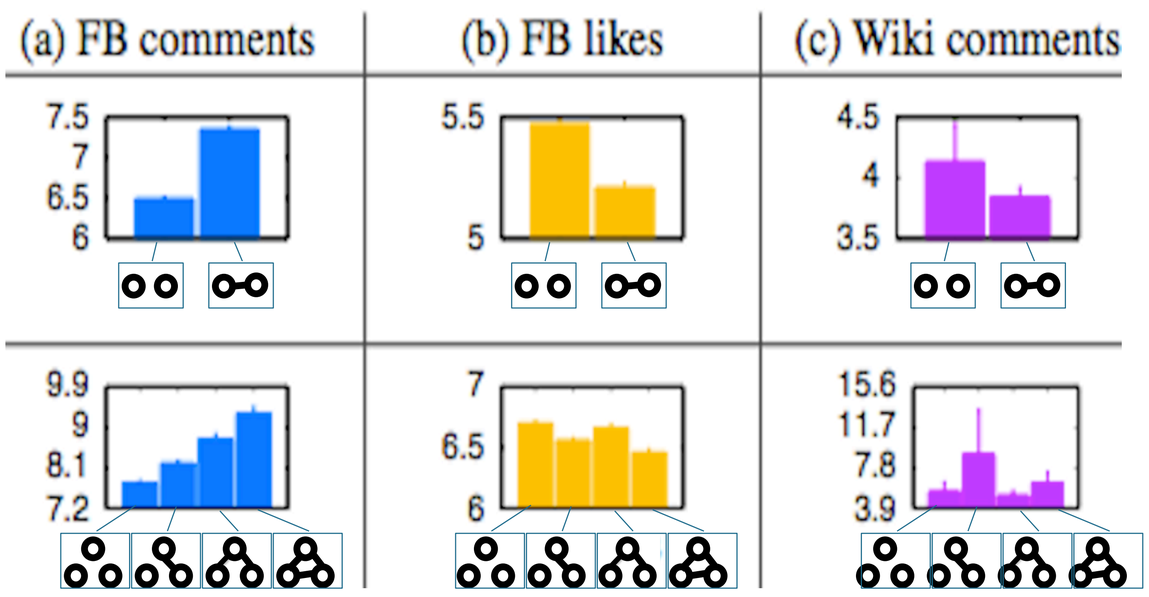
To generate commenters $c_1, c_2, \ldots$ for thread: start with c_0, c_1, each with *weight* $w(c_j) = 1$. Parameters $\alpha \geq 1$ and $\beta \geq 0$.

(i) *Choose commenter c_{j+1} after $c_1, \ldots, c_j$ chosen.* Choose pre-existing thread participant (different from c_j) with probability proportional to the weights. Pre-existing participant $c \neq c_j$ is chosen with probability $w(c)/(\beta + \sum_{c' \neq c_j} w(c'))$, whereas a new participant c_{j+1} is introduced into the thread with probability $\beta/(\beta + \sum_{c' \neq c_j} w(c'))$.

(ii) *Update weights.* If participant c_{j+1} chosen in (i) is a re-entrant, we define $w(c_{j+1}) \leftarrow \alpha w_j(c_{j+1})$, and leave all other weights unchanged. If c_{j+1} is new, we set $w(c_{j+1}) \leftarrow 1$ and for all other pre-existing participants $c \neq c_{j+1}$ we reduce their weights via $w(c) \leftarrow w_j(c)/\alpha$.

SOCIAL, 1,2,… SEQUENTIAL, & TEMPORAL } **features outperform textual features at predicting length and re-entry.**

Social connectedness of first participants



1,2,… Arrival sequence

Chance ID code 1 re-enters: **26.7**% vs. **5.6**%
(For 1,0,1,0,0,… this chance is **47.5**%.)

Top features for Facebook length prediction

Features	Area under ROC curve
Text-only baseline (icon: "**Tt**") ("comment", "agree", "anybody", etc., selected by text regression)	.529
Temporal (time 'til 5th comment)	.695
1,2,… +arrival sequence (# uniq in first 5)	.705
"Tt" +number of words (in 5th comment)	.714
+temporal (time 'til 3rd comment)	.721
"Tt" +contains "?" (in 5th comment)	.7256
1,2,… +arrival seq (ID code of 5th commenter)	.7258
1,2,… +arrival seq (# uniq in first 4)	**.7260**

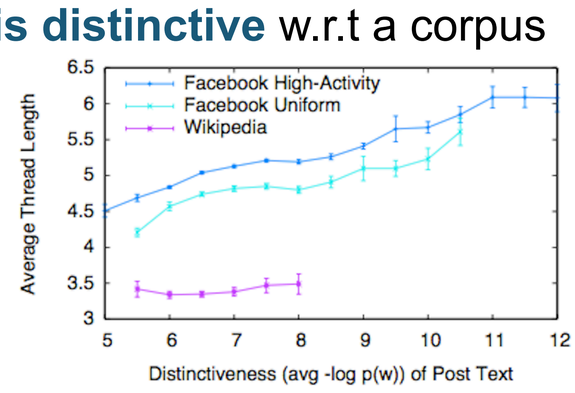
feature-selection order

Top features for Facebook re-entry prediction:
arrival sequence 1,2,…, then temporal

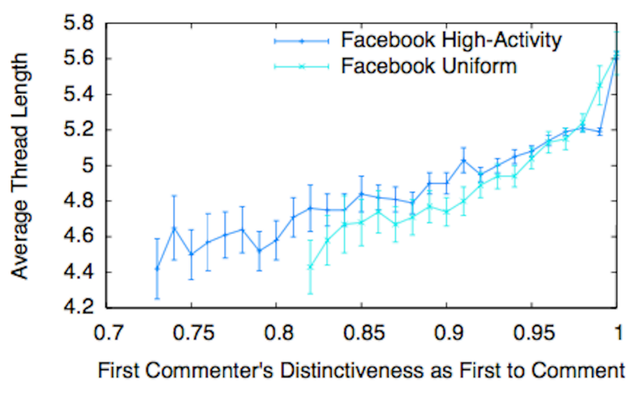
Further directions: distinctiveness

Under certain conditions, **thread length increases when …**

… the post's text is distinctive w.r.t a corpus language model …



… or the 1st commenter is distinctive w.r.t who usually comments on the poster's posts.



Paper/Wikipedia data: http://www.cs.cornell.edu/home/llee/papers/convcuration.home.html