SCRATCH

9/4/14
p. 1. lec 4

Plan: did 'broadcast + social reaction

do we want

structure
SD
Twitter
Youtube: Siersdorfer { length
re-entry

outcomes:
interesting
controversial: Glover, Siersdorfer
vote (wiki) →
speed dates
Resnik / Vinod Prabhakaran
↓ live
↓ email, Enron
email "structure"
single-thread was best per
[illegible]

OUTLINE for Lec 4 on "surface" structure of conversations:

lec notes start on next pg.

discussion can make people adopt more xtreme positions / Schkade et al '07 quoted in Kaltenbrunner's 'structure of political discussions'.

"shape" statistics
- length [Siersdorfer]
- for trees: depth, "bushiness"
[Kumar]
[Gonzalez-Bailon, Kaltenbrunner, Banchs '10]
- dimensions explicitly correspond to conv. aspects (fig 1)
vs. fig 2, eg. discussions
- single measure → render Fig 4.
show txt.
- note anon. users

~~what if you a~~
add participants explicitly
③ email might not even be trees
[Prabhakaran; Rambow '14] - add: drop recipients are actually features!
forwarding = fork.

① Kumar fig 4, section 4.3
avg. # of unique authors / thread length (log-log)
q/a.

② Backstrom et al
distribution of # unique authors
↳ comment length.

add topics explicitly: → quick mention of topic models.
- show Resnik slides (for Nguyen et al. MLJ '14]

only lite in SD - [illegible]
topics that ~~[illegible]~~ recur
topic shifts (show eg. from Nguyen et al paper of shifts)
... relate to power → "Power of confidence - polls"
↳ Enron: org-chart.
[Rob

extension today:
Sente tabs
WSDM poster open

9/4/14
lec 4

: "surface-y" perspectives: what do conv. "look like"

Conversations: thread structure ~~patterns~~ (as opposed to discourse (intentional / structure)

In our 1st lecture, we distinguished btwn 2 sets of ~~systems which~~ ~~that~~ ~~make~~ paradigms where lang & social interaction manifest:

- conversation
- broadcast & social effect

We've spent the last two lectures ~~on~~ doing some preliminary exploration of broadcast & social interaction, ~~so~~

~~Today~~ Now, I'd like to turn to the conversation paradigm.

- - - -> <repeat>

[will also survey diff. sites & corpora]

Thread structure

Let's assume we're dealing w/ a post & ~~comment~~ reply setting:

["comment" is ambiguous: try to stick to 'post' vs. 'reply']

~~A basic statistic~~:
~~We took a look @~~

A basic statistic: length

~~Analysis~~

Analyzing & mining [Siersdorfer et al. 2014]: Chelaru San Pedro Altingovde Nejdl Fig 7. lots of good overview material

Siersdorfer 2014
Youtube ~~comment~~ vs.
Yahoo News
But, they don't allow replies to replies, ~~so~~ in ~~you lose~~ thread structure
That's ok, it's like FB.
Predict what will receive replies

- Comparing Yahoo! news comments to ~~hacker~~ Youtube (~~notably bad com~~ noted "cesspool")
  show page: "posts" hidden by default ~~comm~~ "replies" hidden by default
  lots of ~~comments~~ posts arriving as we speak!
- Histogram of lengths: as expected; ~~not a lot of~~ (~~rare to get a~~
  ~~table V above it: "unreplied" vs "seeds" (as of replies)~~ .
  rare to get a convo going (>50% posts get no comment)

note highlight: control for politics category - still Youtube threads are shorter

~~easier to make new~~ post

affordances. hackernews vs. slashdot, Rosé paper on affordances: SD vs. LegSim.
easier to make new post
"easier" to make new ~~com~~ reply
Elsner paper

- what does a long thread correspond to?
  Fig 9: diff. sign of deriv: for how 'good' initiator of long thread is.

likes vs # comts.

~~For Yahoo, is it~~
Does Yahoo show that ~~more~~ more length = more interest => more ⊕ ratings? (i.e., ⊕ bias combined w/ ~~attention~~ attention?)?

No: Youtube it's the other way. (flame war)

q: really need error bars on fig 9. esp since right-hand bars have little data.

random:

viz of reddit.

Now, what if we consider conversations as trees (explicitly, or, via thread induction, like what Elsner & Charniak '10 did)? Then we have another "dimension" besides length:

2 class members were from Brown!

or Wang/Joshi/Rosé

Also [Wang...

Size,
Depth, & other measures of tree "shape"

Kumar, Mahdian, McGlohon KDD 2010, Dynamics of Conversations.

~~Twitter~~

UseNet groups; Yahoo groups. } graphs for Usenet (others are qualitatively similar)

* Twitter
→ important conv. source we haven't talked about yet.

q: "why Usenet? <which is email, btw>
scan up to part before §3:
"While Usenet is declining in pop, ... public, easy to crawl, obv. thread structure ... some groups still active". This is the rationale

remark about "footnotes in reaction to reviewers". e.x.: "Recall that x is not y". A social interaction!

q: about corpus of reviewer-author "interaction"?
(relate to arXiv: diff. versions; laTeX source available)

~~[illegible]~~

Fig 1b: empirical size vs. depth "fits" a √ law (given the exponent value in the key).

⇒ conversation trees are "deeper" than a "rich-get-richer" would predict (log-depth)
(according to highlighted text)

* counter-claim: the interfaces for these settings don't rank by branching, or etc. so an 'attention bias' towards the rich-get-richer wouldn't happen by affordance.
→ A/B test w/ HackerNews vs. Slashdot? See Wang/Joshi/Rose': SD vs. LegSim = ~~stream~~ reply "stream"

. ~~you~~ as they say, you can hide an elephant in a log-log plot. (checked w/ my signing in revealing my clever login).

Fig ~~1~~ 3: do diff. levels have diff. or same branching?
Lines slope down b/c you expect fewer things to have large branching.
~~But~~ ~~I~~ Point is diff. btwn lines: higher levels (towards tree top) have more branches.

reminder:
... all of this: about what these trees look like.

~~Depth & other measures of tree "shape"~~

Gonzalez-Bailon, Kaltenbrunner, Banchs: - "structure of political ..." - on slashdot!
(for structure w/out partiipants explicitly labeled

abused the term "branching factor" a # of times. (cf. "branching")

max ~~#~~ branching <scratch>

h-index: ~~a~~ 2 measures of depth: "weight in controversy of certain comments."
[engagement of most active users; not relevant for lec. outline]

~~# of layers~~ x ⇒ x layers w/ @ least x comments in that layer

Fig 1: ~~again~~ start w/ theory about (useful discussions), Fig 1.
Vertical ~~access~~ axis: "argumentation level", is practically "literally" depth of the tree
better discussions have participants involved. (norm'd: can't have negative depth!)
~~repr~~
horiz: democratic discussions involves more people
... ⇒ more people pile in @ a particular post, so correlate it w/ branching
[later, also explicit consider # of unique participants in the tree]
↑ in paper

Scan thru: note highlights on "why SlashDot" (remark: "these people really like Slashdot")
"had enough time to evolve & consolidate, overcoming the problems associated to spam or misbehavior and proving its robustness"
: also about handling of anon posts ("Anonymous Coward" is not a single person.)

- don't remove - would break the tree structures.

Fig 2: examples of 2x2 plane (presumably real slashdot trees)
- highlight: quadrant I is the "good" one.

Fig 3: plot political v. non-political on that plane.
"intersection" of dotted lines: mean width & depth.
Centroid shows the political discussions do indeed "look diff" - in Quad I - from non-political

(Fig 4: divide by category. Not in color, unfortunately.
linux: near center
games: very low participation (Quad III): ~~fewer people in~~ less active category?)

An h-index measure, to try to summarize shape in a single #.
- see highlighted text, under fig 4.

also, here's translation ...
(note layers w/ big branching don't have to be consecutive)

Now add: participant ID to ~~there~~ conversation structure

~~Kumar~~

Kumar et al again, fig 4 (and section 4.3):

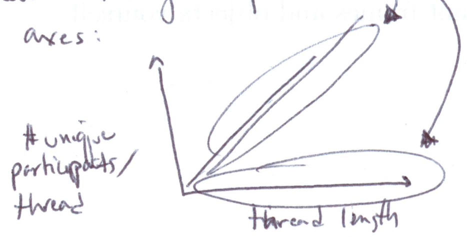
On a log-log, we see that the avg # of unique authors (red) vs. size: a polynomial relationship.

But, what about the distribution of # of unique authors?

(..., cont ~~of "what" is a long thread~~)

< Backstrom, Kleinberg, Lee, Danescu-Niculescu-Mizil WSDM '13 >: (display the WSDM poster)

- guestbook vs. ~~conv~~ involved.
  show heat map on poster:
  axes:



- Hy's law? not in dists (Dyn. of Conv. paper) is avg

SP: dealing w/ avg comments

argue that width is good proxy for # of pyschwilad.

→ counting says this my not be true.

bushiness: h-index.

→ handle of anon.

matters b/c you may want to rate posts by type of thread, not just length of thread
- if it's a thread someone's likely to re-enter, keep them apprised.

- Also since it came up before, note that social-connectedness of 1st few posters has interesting effects on the length of a reply thread vs. a "like" thread

Prabhakaran, Enron
"ca ~~add~~
(add/remove is not for the thread big classify!

- in the setting of email, you may not even have trees "because of" addition of ~~authors/s~~ people's identities

Enron data set (post ref)

forwarding vs. replying <need to clear out the "prior email" text or make sure you know who wrote which parts>
fork

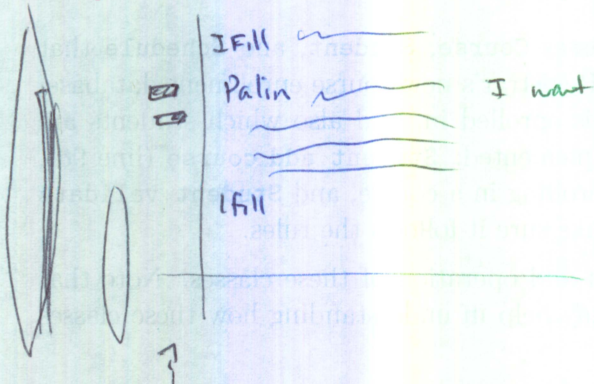
dropping/adding recipients
power relationships known. → (actually used as features in Prabhakaran; Rambow '14)

Note also the UMass work on the North Carolinian corpus (Freedom of Info request?)

Note also the W3C data / Glasgow.

Just talked about adding recipients, what about now adding content to the structure of conv?

<Nguyen, Boyd-Graber, Resnik, Cai, Midberry & Wang MLJ '~~13~~ 14> 95(3): 381-421
(using Philip's slides @ the wksp on Lang & Soc Sci)



I want to talk about, again, my record on energy

Debate
- real-time,
f2f
conversation.

use topic modeling <post ref> to show which topics are active when.
Some recur throughout, some only for a short time.

when does the topic shift?
~~See what~~ From text, Palin is abruptly shifting to a diff. topic.
Ifill tries to get back to other topic.

next slide: ~~same~~ @ moderator & Palin changing topic about the same amt.

next slide: criticism of Ifill for not managing the conversation better

next slide: a diff (set?) of debates.

Prabhakaran & Rambow '14 - topic shifts @ related to poll standings.