facebook

USC Viterbi School of Engineering
Information Sciences Institute

Social Media Analytics

Text Mining for User Profiles and More

Sofus A. Macskassy

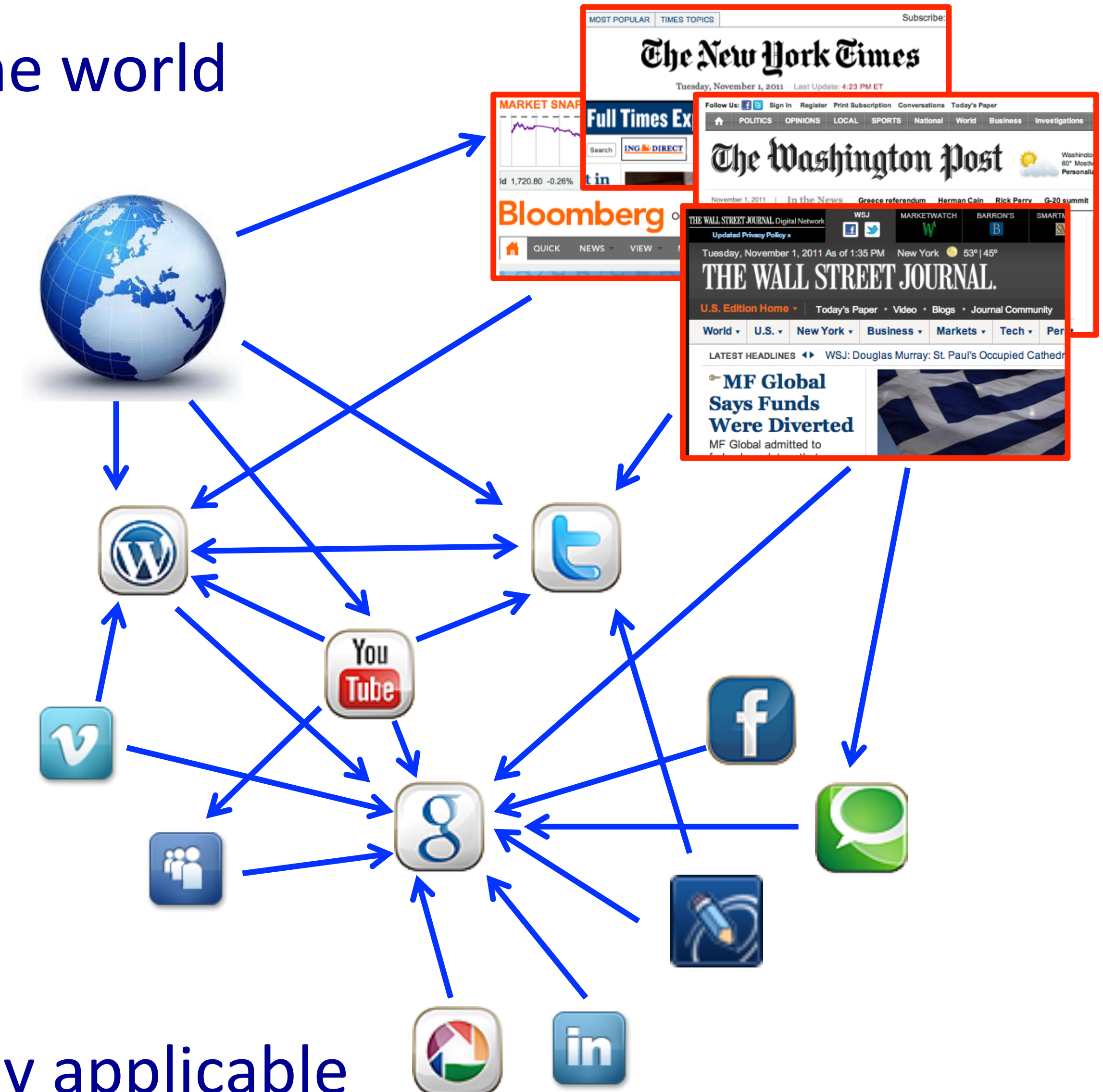
Data Scientist, Facebook

Project Leader, Research Assistant Professor, USC/ISI (on leave)

(sofmac@gmail.com)

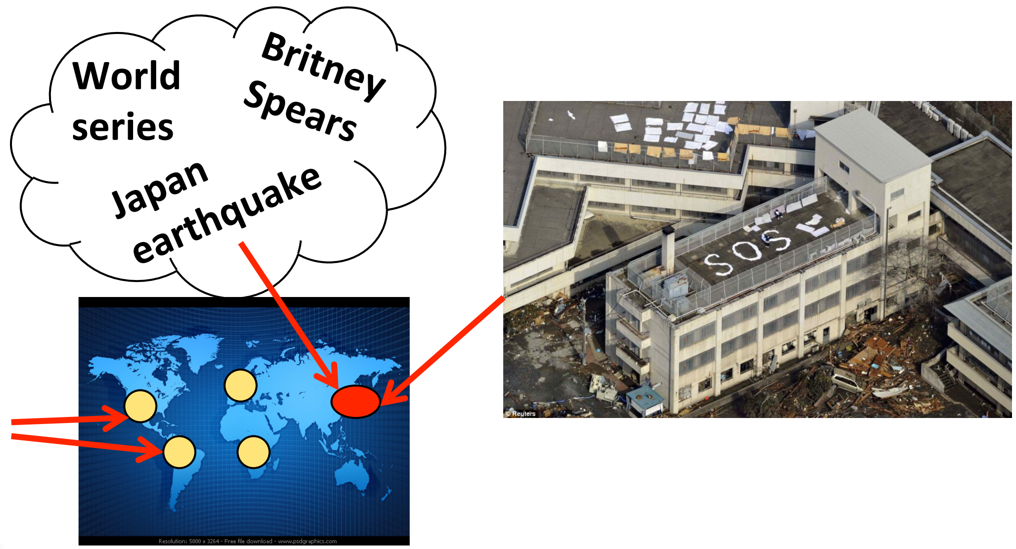
Why Mining Social Media?

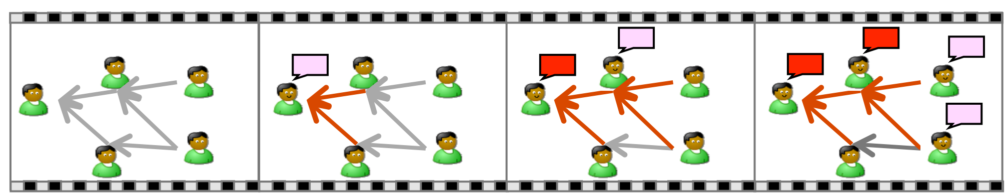
1. It is the pulse of the world
   - People
   - Events
   - News
   - *Real-time*
2. Rich and complex
   - Micro-blogs
   - Blogs
   - News
   - Videos
   - Pictures
   - …
3. Techniques broadly applicable



Making sense of social media

- *What…*
  - Is going on in the world
  - Are people taking about
- *Where…*
  - Is X happening
  - Are people talking about X
- *When…*
  - Did X happen
  - Was X talked about
- *Who…*
  - Is talking about X
  - Is influencing the discussion about X
- *Dynamics*
  - How are things spreading and changing?





Sofus A. Macskassy (sofmac@gmail.com)

Generality of Principles and Techniques

- Complexity of dynamics in content and network
  - Covers a wide variety of disciplines including machine learning, network science, physical systems, social science, economics, game theory, epidemiology, etc.
- Applicable to many problems
  - Better understanding and modeling of social systems
  - Network analysis in financial markets, as defined by social networks, money transactions, etc.
  - Tracking information diffusion by demographics, geospatial, topics, etc.
  - Deep analysis of group-behaviors over time
  - Reasoning with uncertainty about observations and nodes in a network
  - …

Sofus A. Macskassy (sofmac@gmail.com)

Technical Challenges: Many tools still nascent

1. Large and growing amount of data
   - Many sources
   - Noisy and redundant
   - Streaming
2. Data is complex, varied and changing
   - Entity types are increasingly complex
   - Social network keeps changing
   - Resolving topics, entities, etc. across streams very difficult
3. Single-dimensional analytics not enough
   - Must consider network and content both
   - Aggregation of multiple posts/documents for robust analysis
   - Use and integrate different analytic models
4. Human needs to be in the loop
   - Black box systems can suggest but not decide
   - Human intuition needed to guide analytics
   - Interfaces needed to help guide process

My focus is on combining network and content

- Much work in networks and content separately
- Combining them rarely seen in social media mining
  - Machine learning and data mining has statistical relational models (graphical models), but they do not scale well
- Big data problem
  - Dimensions: realtime, volume, heterogeneity
  - The world changes rapidly: what is going on right now
  - → slow models may not be the best
- Answer: start small, then build up
  - Focus on real data from the beginning is key
    - Handle data access, noise, missing data, etc. from the get-go

Agenda

- User profiling from posts
  - Is it possible?
  - What can they be used for
  - When and why to use different representations
- Can we characterize different types of posts?
  - Social dialogues
  - Topics across different types of posts

Recent work

- Blogosphere
  - Semantically tagging of links **[Macskassy 2010, 2011]**
    - Gained new insights into evolution of topics not possible before
  - Demographics **[Michelson and Macskassy, 2011]**
    - Could extract demographic information from blog-posts
    - Demographic clusters and network clusters very different!
    - Could identify demographic characteristics of web-sites based on which bloggers linked to them (in)directly
- Twitter
  - User interests based on tweets **[Michelson and Macskassy, 2010]**
  - Model retweeting behavior **[Michelson and Macskassy, 2011]**
  - Representation matters **[Macskassy, 2012]**
  - What do people talk about?

Twitter

johncrossmirror
User
#Arsenal superb, Cesc magnificent and a brilliant victory. Dunno where the balance is of Arsenal being brilliant and Braga being awful...
"Tweet"
6 minutes ago via web
Off to #Arsenal shortly for Champs League opener with Sporting Braga. Prediction free zone tonight!
Social component
@CallCollymore Sure you already have it, but new Arcade Fire is immense. Have listened to nothing else since I bought it
about 7 hours ago via web in reply to CallCollymore
#Arsenal striker Bendtner insists injury hell coming to an end and his views on Braga: http://tinyurl.com/33feq2f
about 14 hours ago via web
Name John Cross
Location England
Bio Daily Mirror hack and football fanatic who will keep
Social component
140 following
8,639 followers
688 listed
Tweets 6,345
Favorites
Following
View all...

How to represent tweets?

- Problem with tweets
  - They are short (max 140 chars)
  - They contain bad grammar, misspellings, etc.
  - Little context, potentially many words with little overlap
  - Topics are at term level (e.g., Arsenal) not category level (e.g., Football in England)
    - Harder for high level search and clustering → no well-defined topics
- Idea: Leverage entities mentioned
  - From a small sampling of user accounts:
    - ~20% of analyzed Twitter accounts mentioned 50 brands
    - 85% of trending topics are news (likely contain entities)

How to represent user profiles?

Hypothesis

- Analyzing all of users' Tweets can yield their interests
  - Specifically, focus on Named-Entities and the topics they "represent"
  - These topics = the users' topics of interest

Soccer team

English soccer players

Country

User likes topics related to English soccer, international soccer, …

Discover Entities in Tweets

Named Entity Extraction in Tweets

- All caps, all lowercase
- Not often parseable (e.g., POS-tags)
- Look for capitalized (non-stop) words

> #Arsenal winger Walcott: Becks is my England inspiration: http://tinyurl.com/37zyjsc

Disambiguate: Issue

Goal:

- Look up in Wikipedia → get categories → "topics of interest" for user

#Arsenal winger Walcott: Becks is my England inspiration: http://tinyurl.com/37zyjsc

~50 possibilities

Arsenal (Kremlin)
Foochow Arsenal
Arsenal Street
Arsenal F.C. (England)
Arsenal F.C. (Argentina)
Arsenal (Basketball)
Arsenal (Comic)
Arsenal (film)
Arsenal (automobile)
…

~15 possibilities

Walcott, Lincolnshire (England)
Walcott, Iowa (U.S.)
Clyde Walcott
Derek Walcott
Theo Walcott
Mary Walcott
…

3 possibilities

Beck's Brewery
David Beckham
Beckett Scott

Sofus A. Macskassy (sofmac@gmail.com)

Disambiguate: Context

Leverage "context" of Tweet to aid disambiguation

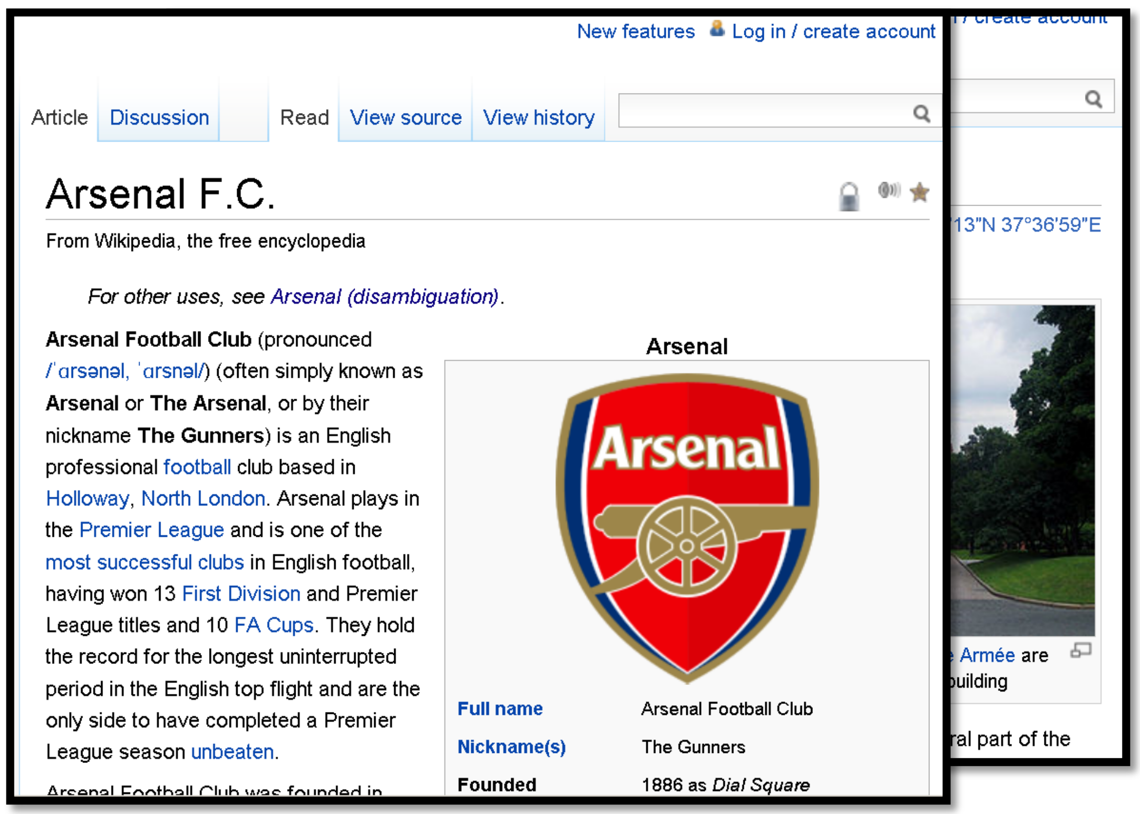
Entity: **Arsenal**

Context: **{winger, Walcott, Becks, England, inspiration, …}**

Language model:
Maximize similarity

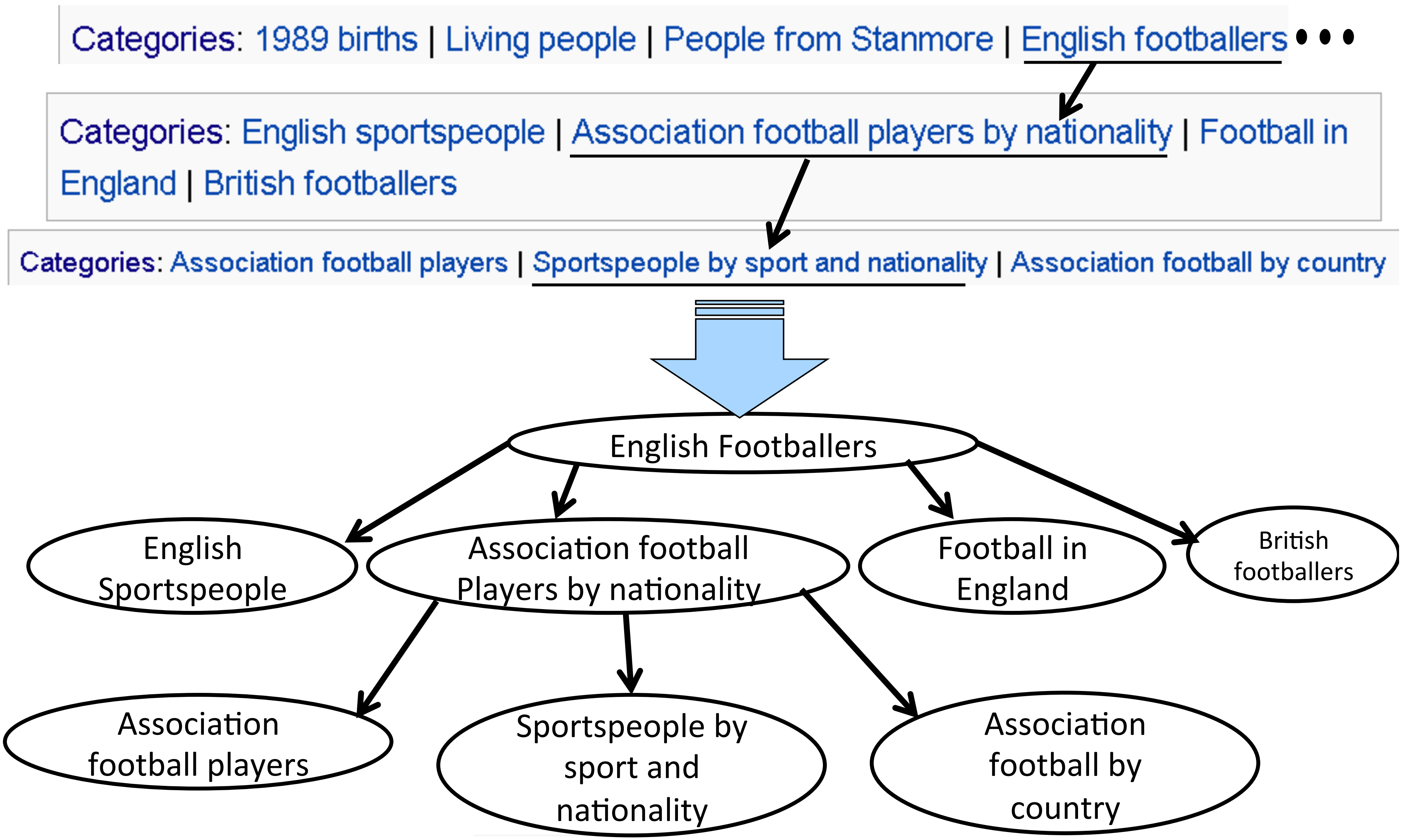
$$\arg\max_{e_i \in E}(C_T \cap C_{e_i})$$

Arsenal = Arsenal F.C.



Sofus A. Macskassy (sofmac@gmail.com)

Folksonomy: "Theo Walcott"



Topic Profiles

After Step 1: Forest of category trees per Tweet

Weight of category c for user u:

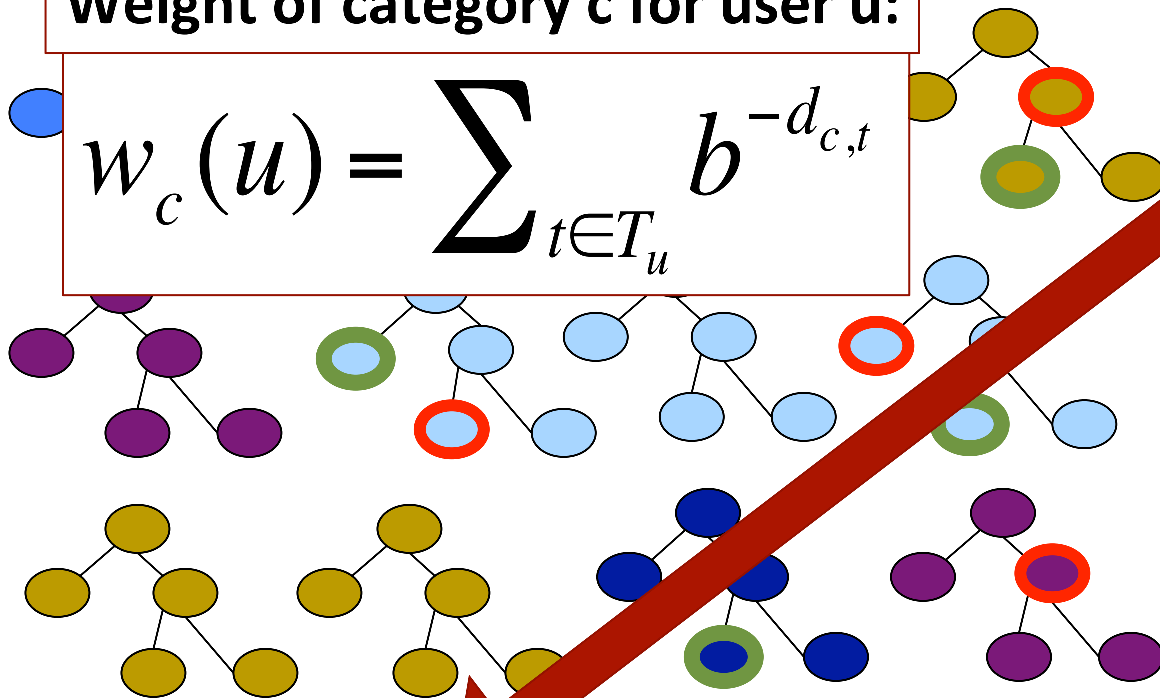
$$w_c(u) = \sum_{t \in T_u} b^{-d_{c,t}}$$

Tweet 1:

Tweet 2:

⋮

Tweet N:



cat62

cat43876

Represent as a vector of (category-id:weight) elements:

{ (cat3,0.5), (cat62,5.4), (cat298,0.2), …, (cat43876,1.7)}

Getting a user profile from generated content

Input Tweets for User

#Arsenal winger Walcott: Becks is my England inspiration: http://tinyurl.com/37zyjsc
Oh, and Senderos's Dad is lovely fella too. Philippe cried in d/r after #Arsenal CL defeat at Liverpool. That shows he cares - wish more did

• • •

Step 1: Discover Categories

Discover Entities in Tweet

Disambiguate Entities

Retrieve Folksonomy Sub-Tree

WIKIPEDIA
The Free Encyclopedia

Step 2: Discover Profile

Generate Topic Profile from Sub-Trees

Topic Profile:
“English Football”
“World Cup”
• • •

Proof-of-concept tiny study: Data

2 “known” Twitter users [know topics of interest]

- johncrossmirror → soccer writer for UK newspaper
- gizmodo → technology blog (loves Apple!)
- Validity for users with specific topics, lots of NE

2 random Twitter users

- Read Tweets by hand to see topics of interest.
- Validity for “random” users

User	# Tweets collected
Johncrossmirror	280
Gizmodo	599
Anonymous1	340
Anonymous2	180

Results: Precision@K

Read posts and examine generated topics

- "Relevant" or "Not Relevant"

Size of Top K	Avg. Precision (± Std. Dev.)
5	0.95 ± 0.10
10	0.90 ± 0.08
25	0.85 ± 0.08

John Cross UK soccer writer	Gizmodo Tech blog (+Apple)	Anonymous1 TV and PAC-10	Anonymous2 Music + Chicago Cubs
ASSOCIATION FOOTBALL PLAYERS	*COMMUNICATION*	*TELEVISION SERIES DEBUTS BY YEAR*	*BASEBALL*
2010 FIFA WORLD CUP PLAYERS	*APPLE INC.*	*2000'S AMERICAN TELEVISION SERIES*	*LIVING PEOPLE*
SPORT IN ENGLAND	*EMBEDDED SYSTEMS*	*ASSOCIATION OF AMERICAN UNIVERSITIES*	*CACTUS LEAGUE*
FOOTBALL IN ENGLAND	COMPANIES ESTABLISHED IN 1976	*AMERICAN TELEVISION PROGRAMMING*	*ALBUMS*
2006 FIFA WORLD CUP PLAYERS	*COMPANIES BASED IN CUPERTINO, CALIFORNIA*	*UNIVERSITIES AND COLLEGES IN THE GREATER LOS ANGELES AREA*	*BASEBALL TEAMS IN CHICAGO, ILLINOIS*
SPORTS TEAMS BY COUNTRY	*TECHNOLOGY*	*PACIFIC-10 CONFERENCE*	*2000'S MUSIC GROUPS*
ASSOCIATION FOOTBALL IN EUROPE	*TELECOMMUNICATIONS*	*SCHOOLS ACCREDITED BY THE WESTERN ASSOCIATION OF SCHOOLS AND COLLEGES*	*SPORTS TEAMS BY SPORT*
ORGANISATIONS BASED IN ENGLAND	*MEDIA TECHNOLOGY*	EDUCATIONAL INSTITUTIONS ESTABLISHED IN 1880	*BASEBALL LEAGUES*
ASSOCIATION FOOTBALL	*COMPUTING*	OLYMPIC INTERNATIONAL BROADCAST CENTRES	*BASEBALL TEAMS*
SPORT IN ENGLAND BY SPORT	*ELECTRONIC HARDWARE*	*NATIONAL ASSOCIATION OF INDEPENDENT COLLEGES AND UNIVERSITIES MEMBERS*	*CHICAGO CUBS*

Discussion

Other topic models (why not LDA?)

- Data is sparse (not many Tweets, they are short)
- Topics are
  - Term level (e.g., Arsenal)
  - Not category level (e.g., Football in England)
  - Harder for high level search and clustering → not well defined topics

Leveraging entities

- ~20% of analyzed Twitter accounts mentioned 50 brands
- 85% of trending topics are news (likely contain entities)

Discussion

Hashtags

- User-given token for search and categorization

Username	Hashtags (ordered)
John Cross	Arsenal, England, wc2010 Spurs, mufc
Gizmodo	iPad, Apple, memoryforever ipadapps, photography
Anonymous1	USC, dadt, glee omgfacts, spoileralert
Anonymous2	Cubs, Nowplaying, Blackhawks Chicago, MLB

- At term level and some are ill-defined
  - Overly specific, difficult to analyze, short life-span, …

  → Hard to use as topics of interest

Agenda

- User profiling from posts
  - ~~Is it possible?~~ (demo at end)
  - What can they be used for
  - When and why to use different representations
- Can we characterize different types of posts?
  - Social dialogues
  - Topics across different types of posts

Question: What makes People Retweet?

- Twitter is an interesting social media broadcast platform
  - Twitterers can broadcast small messages
- Twitter has introduced syntactic constructs to help navigate tweets
  - Retweeting (re-broadcast a messsage):
    - [user_K] I just saw Kevin Bacon at the Kids'R'Us in La Brea!
    - [user_X] RT @user_K I just saw Kevin Bacon at the Kids'R'Us in La Brea!
  - "Messaging"
    - [user_K] @user_X Hey, stop retweeting me!
  - "Topics"
  - [user_K] #KevinBacon He just let Kids'R'Us with a huge doll house!
- The question we ask here is *why* do users retweet?
  - Is it based on things they like?
  - Is it based on who posts?

Hypothesis

- Knowing more about a user's interests and past behavior can help predict his or her future retweets
- We will develop and test three hypotheses
  - Homophily: Users are more likely to retweet information coming from *people who are like them*

    → Similarity of two profiles
  - Topic: Users are more likely to retweet information *they find interesting*

    → Similarity of user profile to tweet
  - Network: Users are more likely to retweet information from people *they were in recent communications with*
- Null hypothesis: Random model

Retweet models

- Null model (General "recency" model)

$$P_{\text{gm}}(x) = c \cdot time(x)^{-\alpha}$$

- Networking

$$P_{\text{network}}(x) = P_{\text{gm}} \cdot \begin{bmatrix} \alpha \cdot P(x \mid recent(x)) + \\ (1-\alpha) * P(x \mid !recent(x)) \end{bmatrix}$$

- Topic

$$P_{\text{topic}}(x) = P_{\text{gm}} \cdot P_{\text{ts}}(x \mid \text{sim}_{\text{topic}}(x,u))$$

- Homophily

$$P_{\text{homophily}}(x) = P_{\text{gm}} \cdot P_{\text{ps}}(x \mid \text{sim}_{\text{homophily}}(x,u))$$

Computing similarity

- We use a standard cosine distance metric to compute similarities

$$sim(\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \cdot \|\mathbf{v}_2\|}$$

Retweet Study: Data

Collected 4 weeks of tweets from ~30K Twitterers

Using geographic-based snowball sampling

- Selected seed ~200 Twitterers in Pakistan and Israel
  - Extracted users mentioned retweet or messages
  - Added users who were (self-reportedly) in Pakistan or Israel
- Increased to ~30K Twitterers in a matter of a week

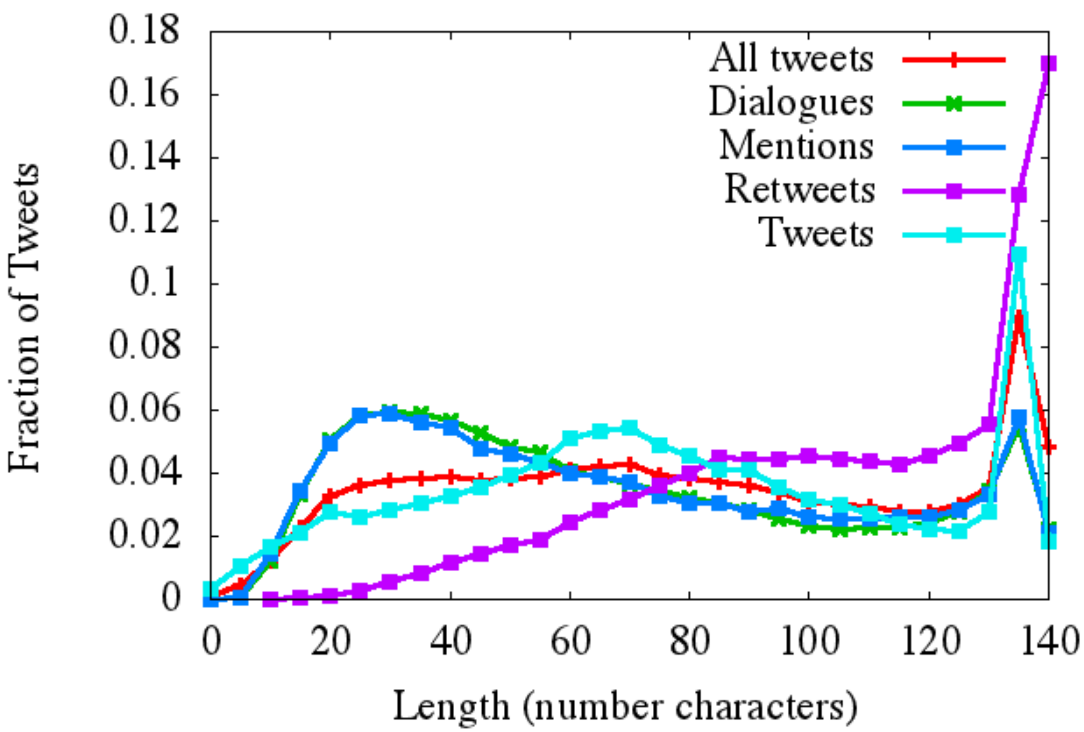
Using tweets from users from 9/20/10 to 10/20/10

- 768K tweets

Retweet Study: Data (cont'd)

Consider only users who had 3+ tweets and 3+ retweets

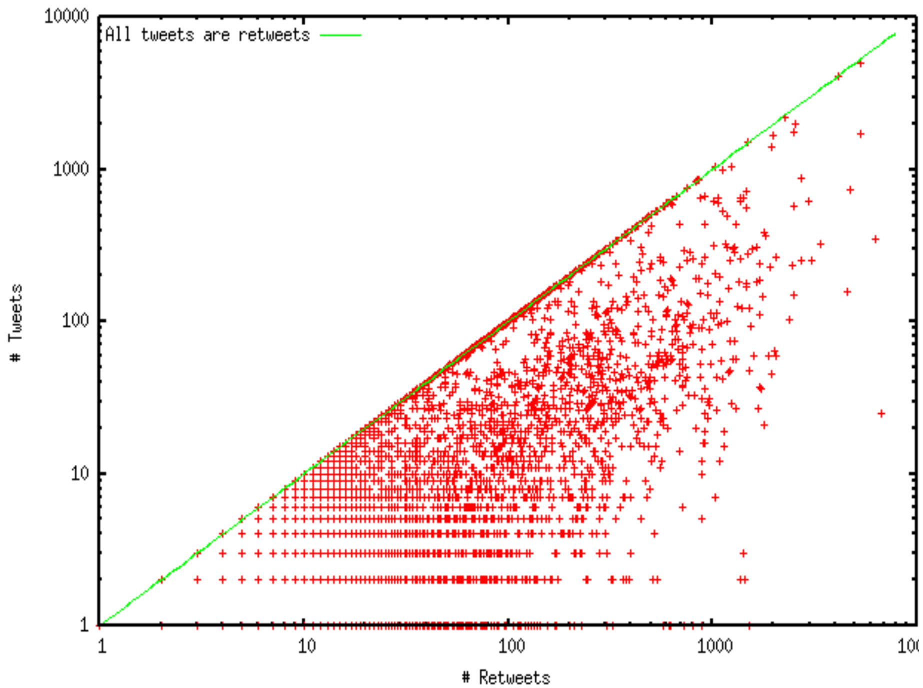
- 482K tweets (**43%** have concepts; **84%** have words)
- 103K retweets (**70%** have concepts; **94%** have words)
- 16K retweets of 1800 users where both original tweeter and retweeter had 3+ tweets and 3+ retweets

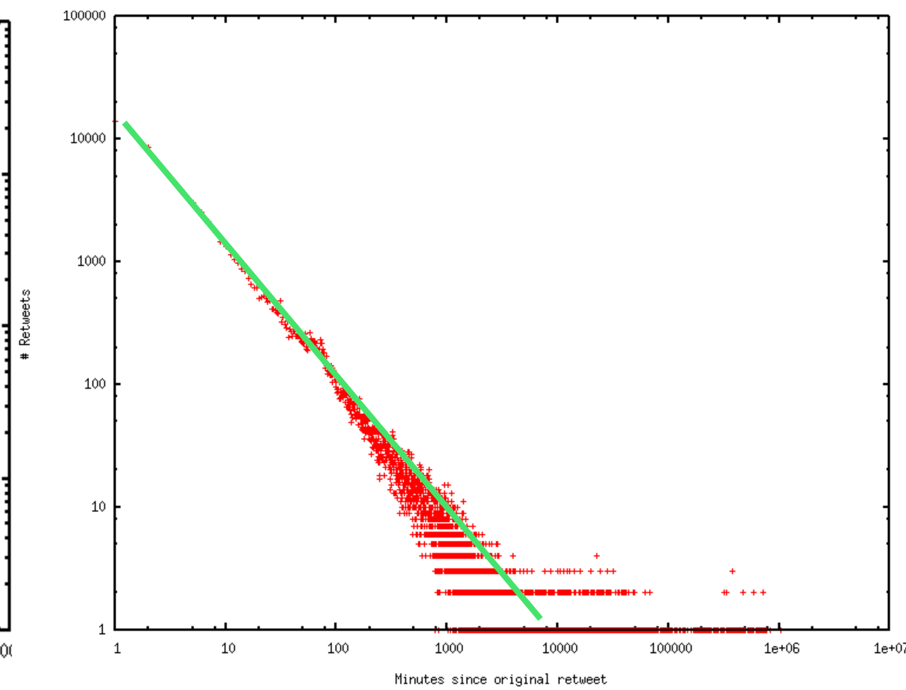


Fitting the models / Evaluation

- We have four models
- We fit the parameters to the data and evaluate which model best fits the data

How much do users retweet? (log-log-scale)



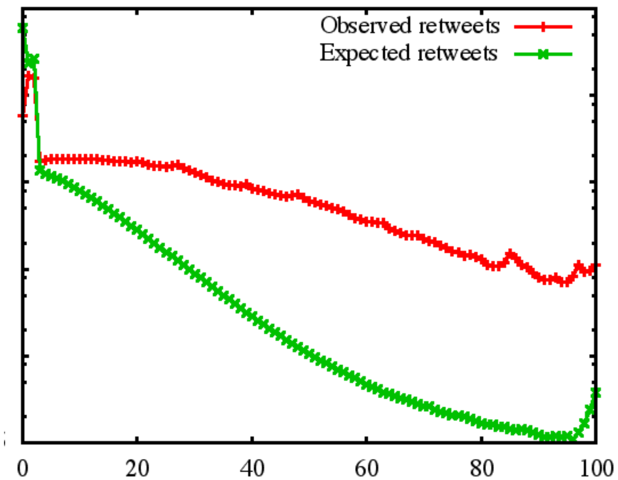


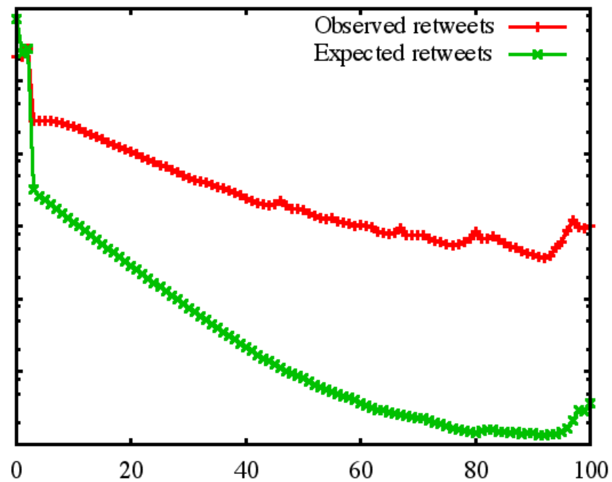
How much do people retweet?
→30% of tweets are retweets
(log-log scale)

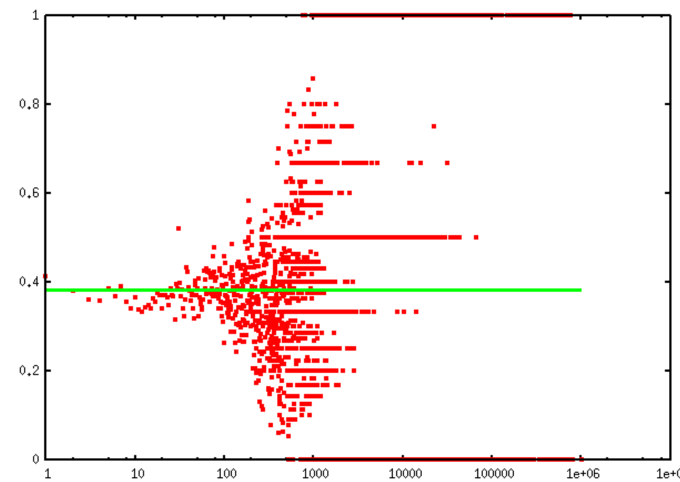
What is the recency of retweets?
Tend to retweet recent tweets
Powerlaw: α=1.15
(log-log scale)

Null Model

Retweet behavior: What does the data tell us?







<u>Homophily:</u>
More likely to retweet someone similar?
(y log-scale)

$$P_{psim}(x \mid \mathrm{sim}_{\mathrm{P}}(x,u))$$

<u>Topic:</u>
More likely to retweet something interesting?
(y log-scale)

$$P_{\mathrm{tsim}}(x \mid \mathrm{sim}_{\mathrm{T}}(x,u))$$

→ Use logodds ratio at given similarity percentile

<u>Network:</u>
38% of retweets to someone recently (<1 day) retweeted
60% of retweets are to someone previously retweeted
(x log-scale)

Retweet models

- Null model (General "recency" model)

$$P_{\mathrm{gm}}(x) = 0.2 \cdot time(x)^{-1.15}$$

- Networking (α =0.38)

$$P_{\mathrm{network}}(x) = P_{\mathrm{gm}} \cdot \begin{bmatrix} \alpha \cdot P(x \mid recent(x)) + \\ (1-\alpha) * P(x \mid !recent(x)) \end{bmatrix}$$

- Topic

$$P_{\mathrm{topic}}(x) = P_{\mathrm{gm}} \cdot P_{\mathrm{ts}}(x \mid \mathrm{sim}_{\mathrm{topic}}(x,u))$$

- Homophily

$$P_{\mathrm{homophily}}(x) = P_{\mathrm{gm}} \cdot P_{\mathrm{ps}}(x \mid \mathrm{sim}_{\mathrm{homophily}}(x,u))$$

Evaluation

- Questions we ask:
  - Globally: Which models fits bets?
    - Might be dominated by prolific users (10% generating 90% of traffic)
  - By user: Which model fits best?
    - Might not get diversity of users
  - By tweet: How often is each model used per user?
- Second, does the representation have an effect?
  - Concepts vs. text vs. hybrid
- Finally: what do the networks look like? Can we not just look at the follower-graph to mine diffusion?

Evaluation using concept representation

Globally

Homophily (45% of all retweets)

By User

Ratio of users explained by each model

Null Model	Network	Homophily	Topic
12%	14%	67%	31%

On average (by tweet), users used the following models

Null Model	Network	Homophily	Topic
11%	26%	37%	26%

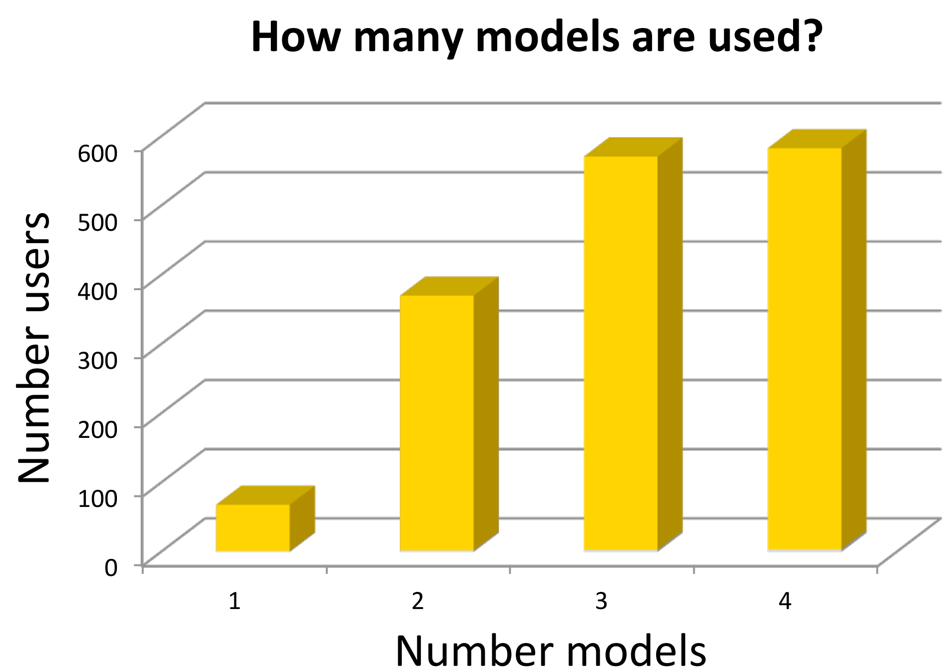
Evaluation using concept representation

By Tweet

On average, users used the following models

Null Model	Network	Homophily	Topic
11%	26%	37%	26%

- **User's behaviors actually best explained by a mixture of models**
- **How many many models needed to explain all of a user's behavior?**



Agenda

- User profiling from posts
  - ~~Is it possible?~~ (demo at end)
  - ~~What can they be used for~~
  - When and why to use different representations
- Can we characterize different types of posts?
  - Social dialogues
  - Topics across different types of posts

What about using a text-based profile?

- We use a standard ***tf.idf*** model, where we consider all tweets (not retweets) from a person as one long document
- We do ***stemming*** and ***stop-listing*** to reduce dimensionality
- We then represent a profile of a user as a standard bag-of-words tf.idf vector of (wordID, wt) elements:

Represent as a vector of (word-id:weight) elements:
{ (wing,0.5), (soccer,5.4), (beck,0.2), …, (uefa,1.7)}

Weights are computed using the Okapi BM25 ranking:

$$w_w(u) = IDF(w) \cdot \frac{f(w,u)\cdot(k_1+1)}{f(w,u)+k_1\cdot\left(1-b+b\cdot\frac{|D_u|}{avgD}\right)}$$

$k_1 = 1.8$
b = 0.75

$$IDF(w) = \log\frac{N_u - n(w)+0.5}{n(w)+0.5}$$

We prune rare words (<5 users use them)

Computing similarity

- We use a standard cosine distance metric to compute similarities

$$sim(\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \cdot \|\mathbf{v}_2\|}$$

- This works equally well when comparing text vectors or category vectors
- Can also get a ***hybrid*** score by taking the average of the text and category similarities

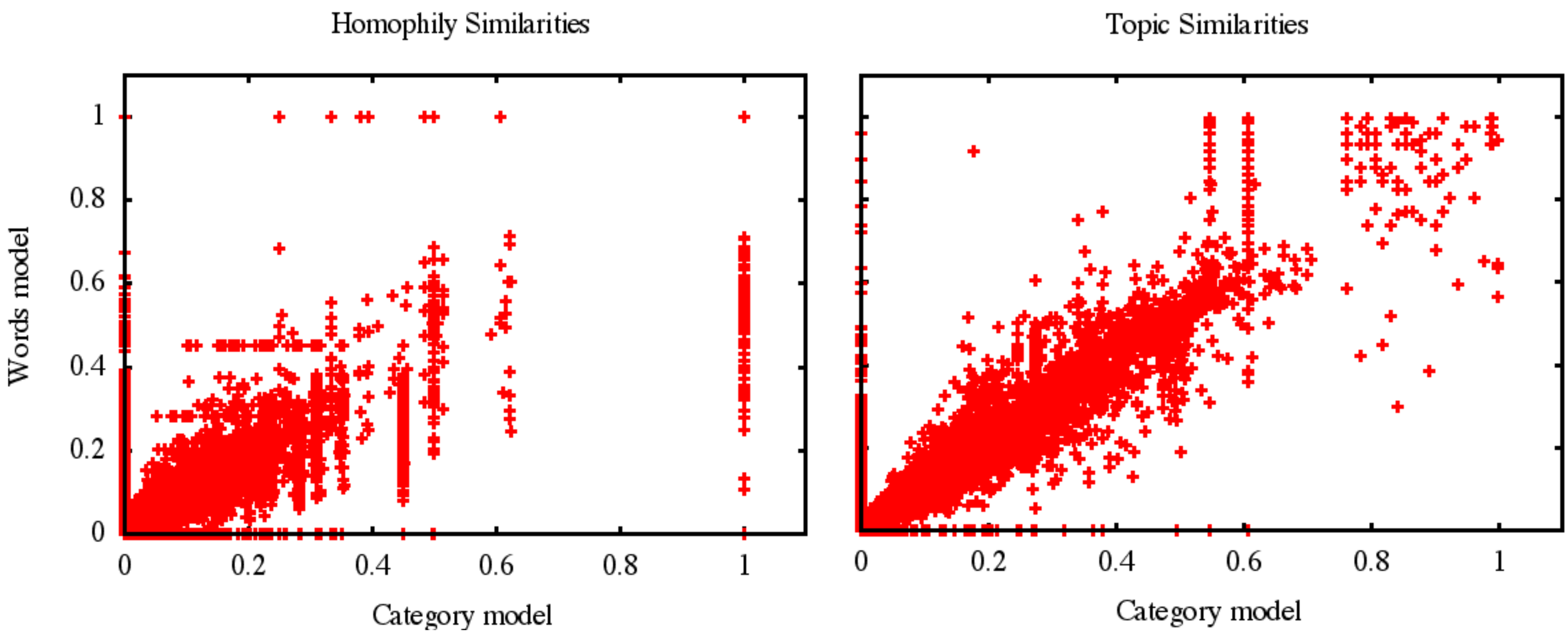
➢ So we can now test our hypotheses across the three models *as well as across 3 representations*

Evaluation: Which representation to use?

- We have three types of representations
  - Concepts
  - Text
  - Hybrid (averaging the above two)
- Does representation have an effect on fit?
  - Some tweets may not have concepts
    - Therefore words might be better
  - Some tweets on same concept may not share words
    - Therefore using concepts might be better
  - A hybrid might be the best of both worlds?

How does text- and category-profiles correlate?

- How do similarity scores correlate between the word and concept representations?



Retweet Study: Data

- Collected 4 weeks of tweets from ~30K Twitterers
- Using geographic-based snowball sampling
- Using tweets from users from 9/20/10 to 10/20/10
  - 768K tweets
- Include users who had 3+ tweets and 3+ retweets
  - 482K tweets (**43%** have concepts; **84%** have words)
  - 103K retweets (**70%** have concepts; **94%** have words)
  - 16K retweets of 1800 users where both original tweeter and retweeter had 3+ tweets and 3+ retweets

Global evaluation across different representations

Concept Models

Model	Wins	Pct
Null	1446	9%
Network	4918	32%
Homophily	**7037**	**45%**
Topic	3183	20%

Text Models

Model	Wins	Pct
Null	946	6%
Network	4976	32%
Homophily	**6486**	**42%**
Topic	3834	**25%**

Hybrid Models

Model	Wins	Pct
Null	1005	6%
Network	5229	34%
Homophily	**6591**	**42%**
Topic	3390	**22%**

By user evaluation across representations

Concept Models

Model	Wins	Pct
Null	203	12%
Network	234	14%
Homophily	**1116**	**67%**
Topic	518	31%

Text Models

Model	Wins	Pct
Null	145	9%
Network	222	13%
Homophily	**1076**	**65%**
Topic	597	**36%**

Hybrid Models

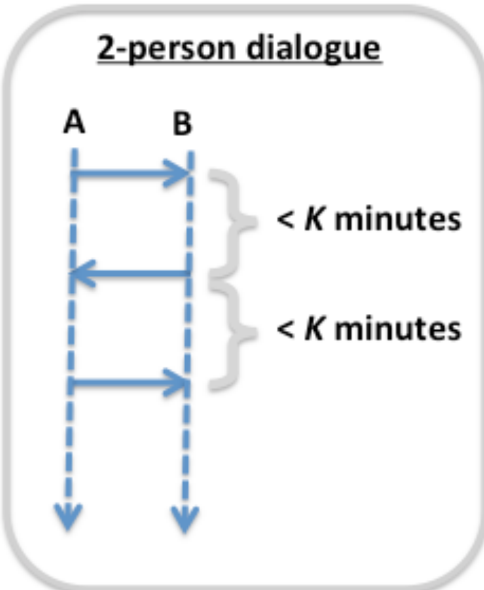
Model	Wins	Pct
Null	145	9%
Network	232	14%
Homophily	**1143**	**69%**
Topic	529	**32%**

- Text works better for Topic model because individual tweets have less data
- Concepts work better for Homophily-model because high-level concepts better captures a user's interest
- Hybrid representation had *best overall* performance for Homophily model

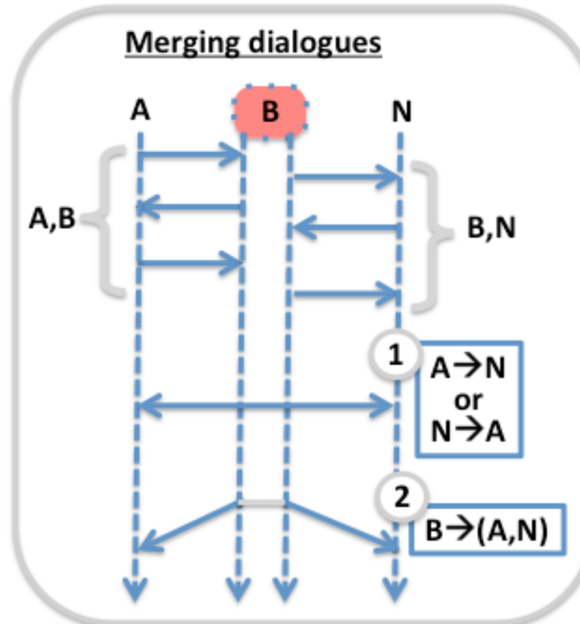
Agenda

- ~~User profiling from posts~~
  - ~~Is it possible?~~ (demo at end)
  - ~~What can they be used for~~
  - ~~When and why to use different representations~~
- Can we characterize different types of posts?
  - Social dialogues
  - Topics across different types of posts

Extracting Dialogues

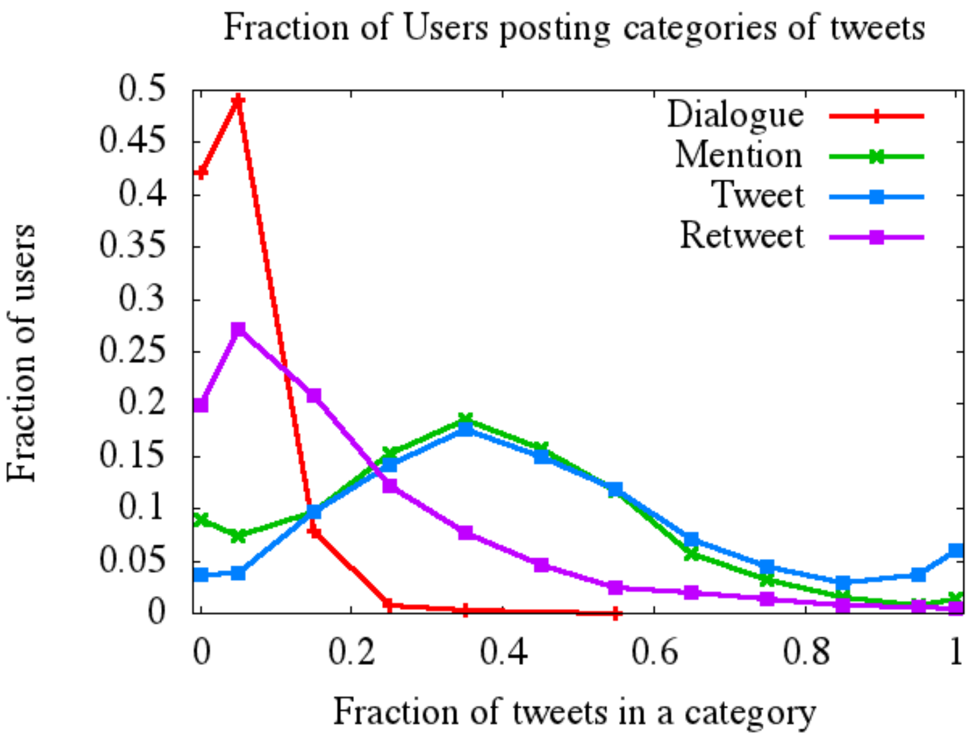


Time	User	Tweet
23:56	**User1**	**@User2** why don't you get a car my friends
00:00	**User2**	**@User1** cause my cars transmission blew before i
00:01	**User1**	**@User2** ohh and when you come back we must g
00:03	**User2**	**@User1** DEFINITELY im going there and in an o
00:09	**User1**	**@User2** do you not have a in n out to??
00:16	**User2**	**@User1** no we do but its hella far :(i come back i
00:19	**User1**	**@User2** my birthday!! I'll drive??
00:23	**User2**	**@User1** im sooo down..my parents wanna get me



ID	Time	User	Tweet
(A)	18:55	**User1**	**@User7** MQM is THE MAFIA! The organi
(A)	18:57	**User1**	**@User2** bro, please stop misstating me. I lo
(A)	19:00	**User2**	**@User1** Mafia of MQM makes 70% of Kar
(B)	19:01	**User3**	**@User4 @User5 @User2** v hope 4 a politi
(A)	19:04	**User1**	**@User2** 70% of Karachi is MQM? Really?
(C)	19:05	**User2**	**@User6 @User1 @User5 @User3** MQM
(C)	19:06	**User2**	**@User1** ok then app battaa do... Laikin baat
(C)	19:09	**User4**	**@User3 @User5 @User2** :) No Maseeha o
(C)	19:10	**User2**	**@User5 @User6 @User1 @User3** wrong,
(C)	19:11	**User2**	**@User4 @User3 @User5** no doubt about t
(C)	19:13	**User1**	**@User2** :) Your 'facts' tell me this discussi

How do people spend their time?

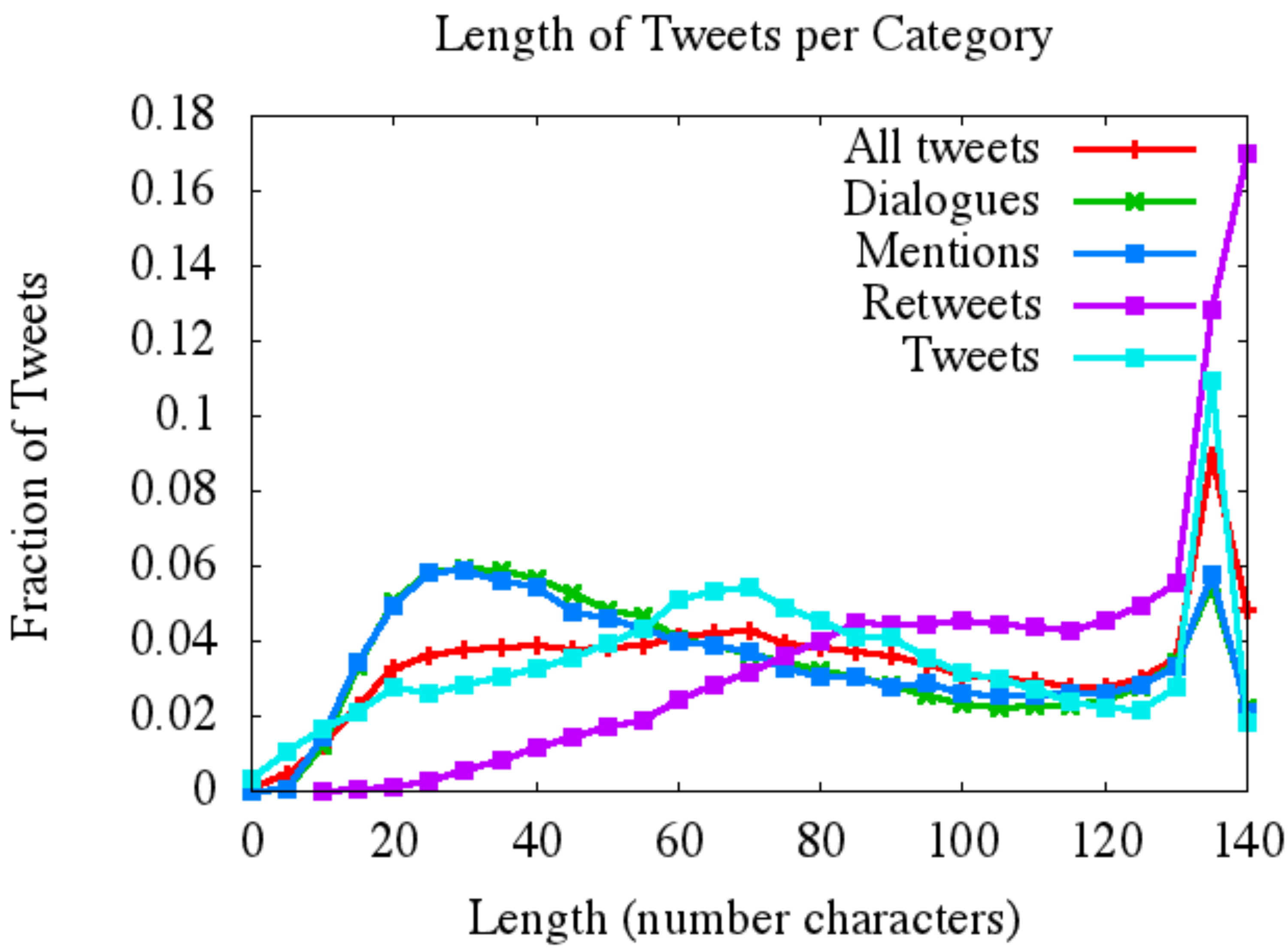


Tweet Category	Number Tweets	Overall Ratio
Dialogue	66,812	0.13
Retweet	93,319	0.19
Mention	154,177	0.31
Tweet	183,748	0.37
Total	498,056	1.00
Conversion	20,155	0.12

Size	Number	Ratio	Avg. Num. Tweets
2	18,619	92.37%	4.9
3	1,232	6.11%	8.5
4	181	0.90%	12.7
5	83	0.41%	19.4
6	27	0.13%	36.5
> 6	13	0.07%	> 60

Activity profiles of users: what fraction of users spend 0% through 100% of their time posting each type of tweet?

Tweet characteristics by class



How is user attention split across friends?

Entropy for user ***u***:

<u>Over dialogues:</u>

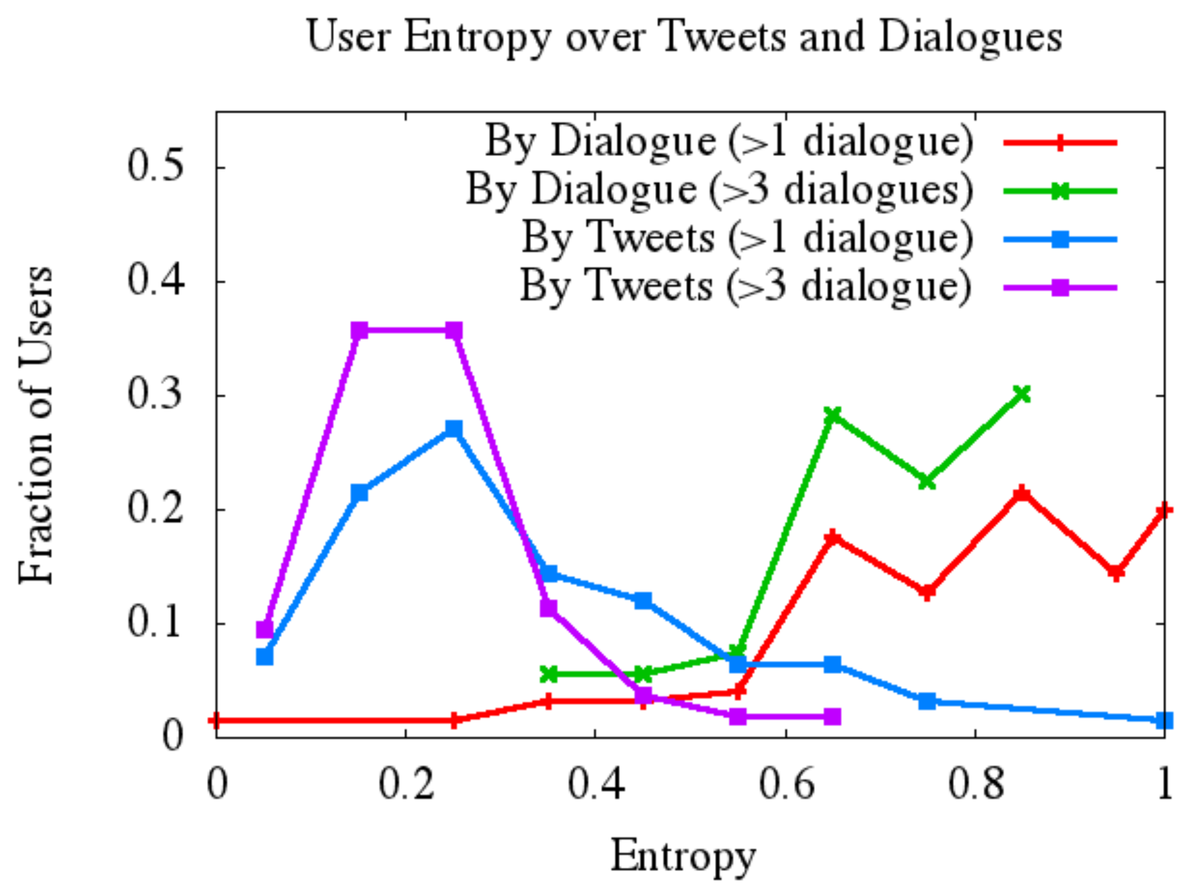
$$-\sum_{n \in N^u} \begin{matrix} d_{u,n} \log(d_{u,n}) + \\ (1-d_{u,n}) * \log(1-d_{u,n}) \end{matrix}$$

$\boldsymbol{d_{u,n}}$= fraction of ***u***'s dialogues which include ***n***

<u>Over tweets:</u>

$$-\sum_{n \in N^u} \begin{matrix} r_{u,n} \log(r_{u,n}) + \\ (1-r_{u,n}) * \log(1-r_{u,n}) \end{matrix}$$

$\boldsymbol{r_{u,n}}$= fraction of ***u***'s tweets which include ***n***



<u>Takeaway:</u>

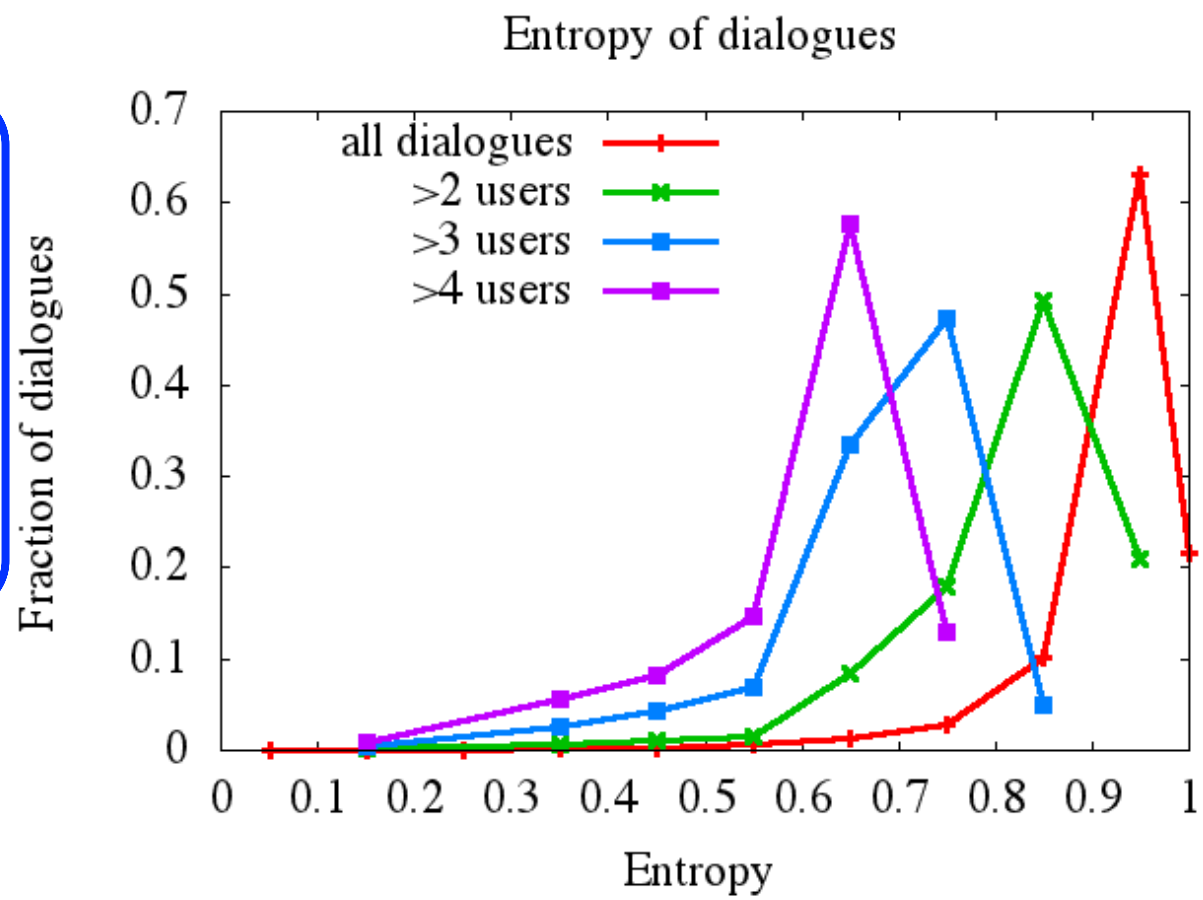
Dialogues cover many people, but overall users interact directly with only a few

How many people *really* participate?

Entropy for dialogue ***D***:

$$-\sum_{u \in D} \begin{matrix} r_{u,D} \log(r_{u,D}) + \\ (1 - r_{u,D}) * \log(1 - r_{u,D}) \end{matrix}$$

$r_{u,D}$= fraction of ***D***'s tweets which originated with ***u***



Takeaway:

The more users participating in a dialogue, the more likely that a few people dominate the discussion

Sofus A. Macskassy (sofmac@gmail.com)

Agenda

- ~~User profiling from posts~~
  - ~~Is it possible?~~ (demo at end)
  - ~~What can they be used for~~
  - ~~When and why to use different representations~~
- Can we characterize different types of posts?
  - ~~Social dialogues~~
  - Topics across different types of posts

What are people talking about?

Different types of posting behaviors

- Retweeting: *information diffusion*
  [Macskassy and Michelson 2011; Macskassy 2012]

- Social dialogues/chat: *social networking*
  [Macskassy 2012]

- General: *Broadcast* tweets to all followers

How do they differ in terms of topics?

Topic Study: Data

- Using tweets from users from 9/20/10 to 10/20/10
  - 768K tweets
- Categorize tweets into dialogues, retweets and other
  - Dialogues: 108K tweets
  - Retweets: 116K tweets
  - Other (general): 429K tweets

Sofus A. Macskassy (sofmac@gmail.com)

Topic Study: Methodology

- Used LDA topic-model
  - Identified 150 topics per tweet category
  - 450 total topics
- Manually labeled 450 topics
  - 80 emergent categories
- Counted how often each category was present in each tweet category

Topics for each kind of behavior

Dialogues (108K)

Categories (37)	Pct
Small-talk	45.8%
Daily life	16.4%
Twitter/FB/...	12.3%
Justin Bieber	10.3%
School	6.9%
Music	5.3%
Complaining	4.8%
Sports	4.1%
Work	3.7%
TV	3.8%
...	

General tweets (429K)

Categories (43)	Pct
Small-talk	17.5%
Breaking news (event)	8.1%
Complaining	6.8%
Technology	6.7%
News	6.2%
Politics	5.4%
Twitter/FB/...	5.3%
Sports	4.7%
Daily life	4.1%
Religion	4.0%
...	

Retweets (116K)

Categories (45)	Pct
News	14.4%
Small-talk	9.3%
Breaking news (event)	8.3%
Politics	7.8%
Technology	6.7%
Justin Bieber	5.9%
Religion	4.9%
Pakistan	4.5%
Music	4.4%
Sports	4.2%
...	

Color Legend:

- Social / Personal topic: (life) "I didn't have time to buy groceries today."
- Public topic: (breaking news): "The flood has now destroyed 100 homes"
- Smalltalk: "Hey, it's been a while. How are you doing?"

Agenda

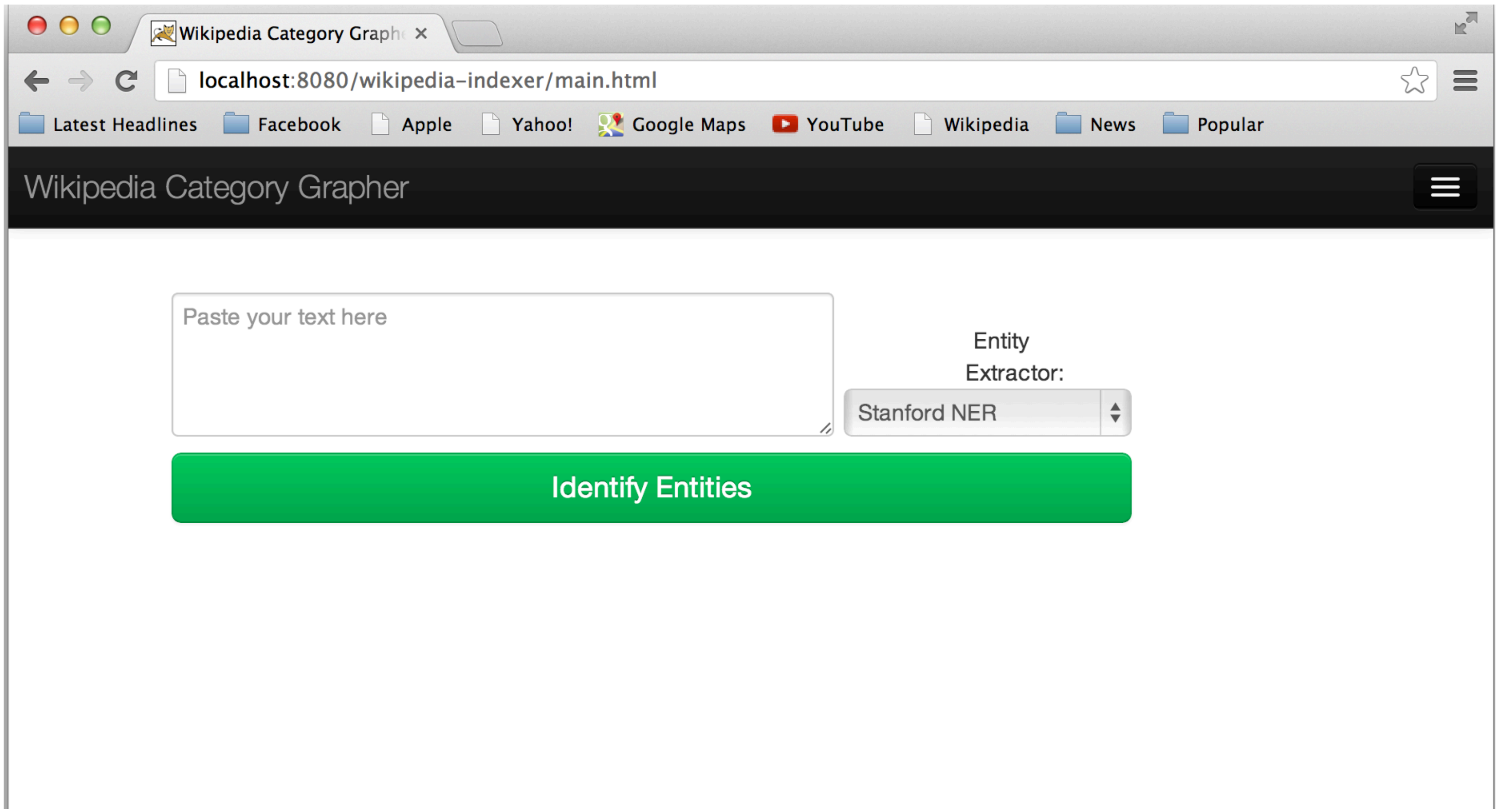
- ~~User profiling from posts~~
  - ~~Is it possible?~~ (demo at end)
  - ~~What can they be used for~~
  - ~~When and why to use different representations~~
- ~~Can we characterize different types of posts?~~
  - ~~Social dialogues~~
  - ~~Topics across different types of posts~~

Summary

- Social media analytics is a rich domain for many different mining and learning technologies
  - Technologies developed here are broadly applicable
  - While joint models likely better, improvements in core technologies crucial to move field forward
- Today I focused primarily on text mining and its uses
  - User profiling and underlying representations
    - Text is good when analyzing single tweets
    - Higher-level mapping more salient for aggregate analysis
  - Explaining retweeting behaviors
  - Homophily rules the day
  - Clear difference in topics for different categories of posts
- Still need a lot of work!

Demo

https://github.com/InformationIntegrationGroup/EntityExplorer



Sofus A. Macskassy (sofmac@gmail.com)

https://github.com/InformationIntegrationGroup/EntityExplorer

Latest Headlines Facebook Apple Yahoo! Google Maps YouTube Wikipedia News Popular

AA AI has a nice Symposium on Data Mining and Text Analytics. Experts in AI and NLP are here.

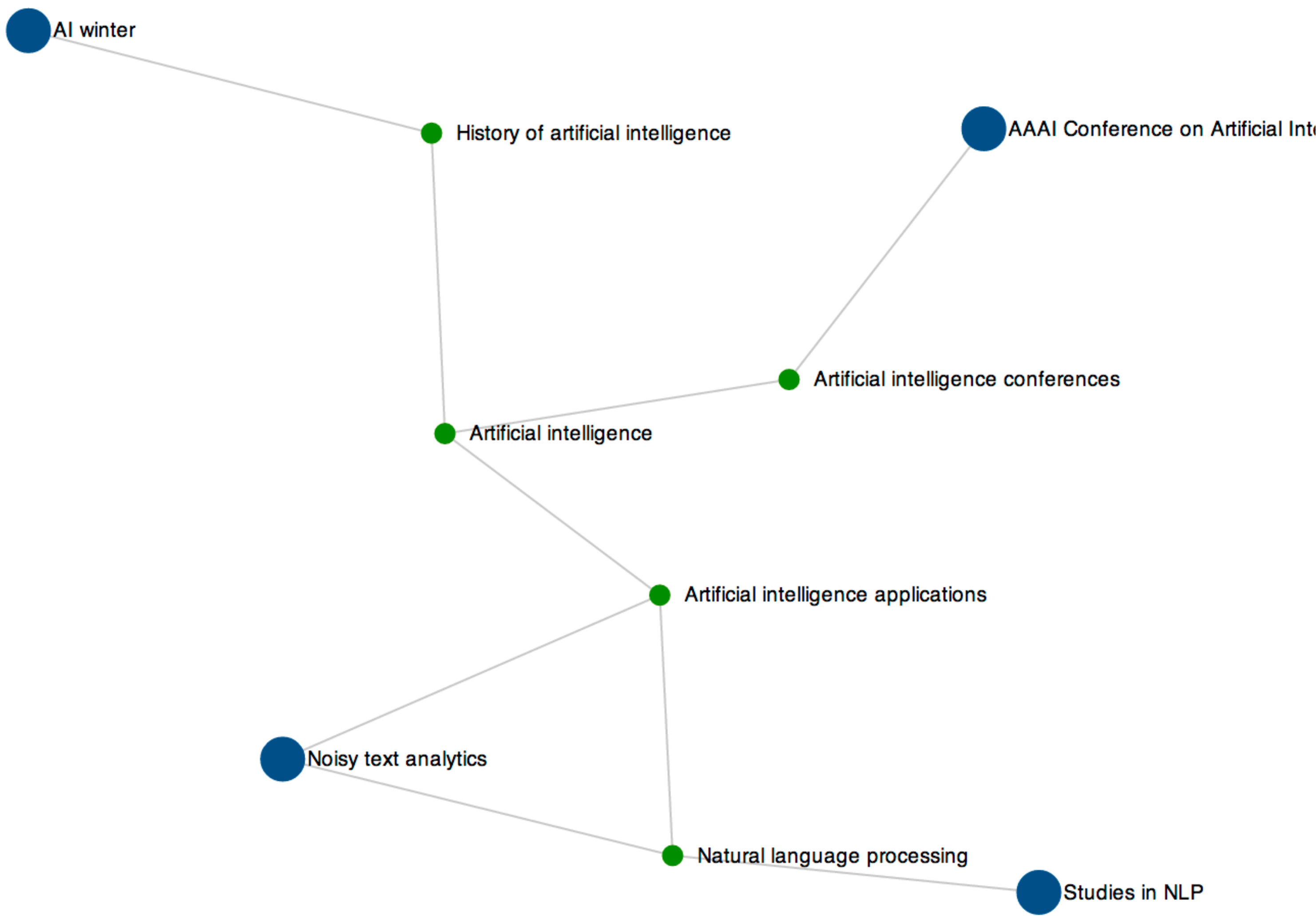
Entity Extractor: Simple Capitalization

Identify Entities

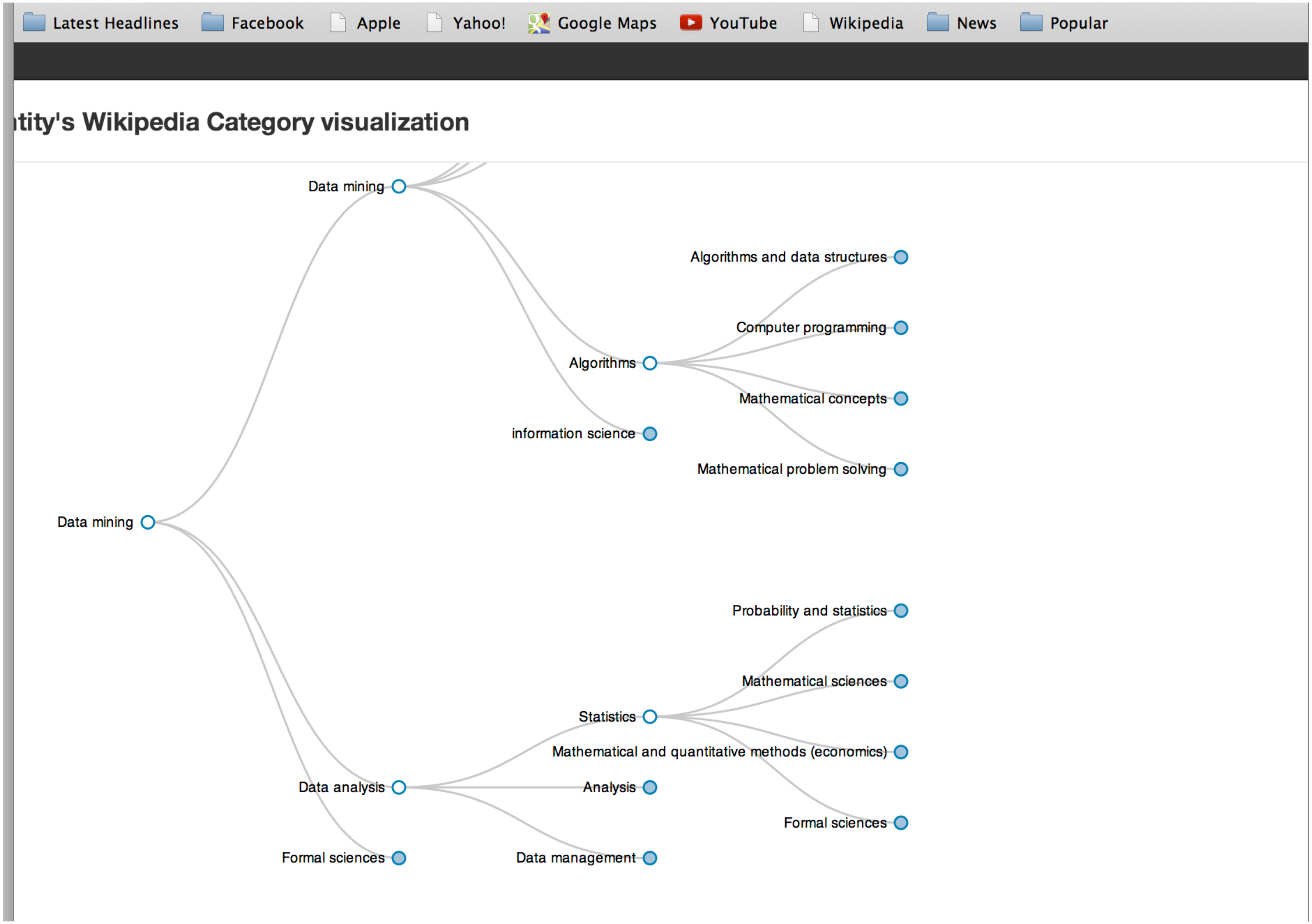
7 entities found

Wikipedia Entity Name	Use in Graph	Wikipedia Categories
Noisy text analytics	☑	View Categories
Data mining	☑	View Categories
AAAI Conference on Artificial Intelligence	☑	View Categories
Studies in NLP	☑	View Categories
Experts-Exchange	☑	View Categories
AI winter	☑	View Categories
Symposium	☑	View Categories

https://github.com/InformationIntegrationGroup/EntityExplorer



https://github.com/InformationIntegrationGroup/EntityExplorer



Thank you

Sofus Macskassy