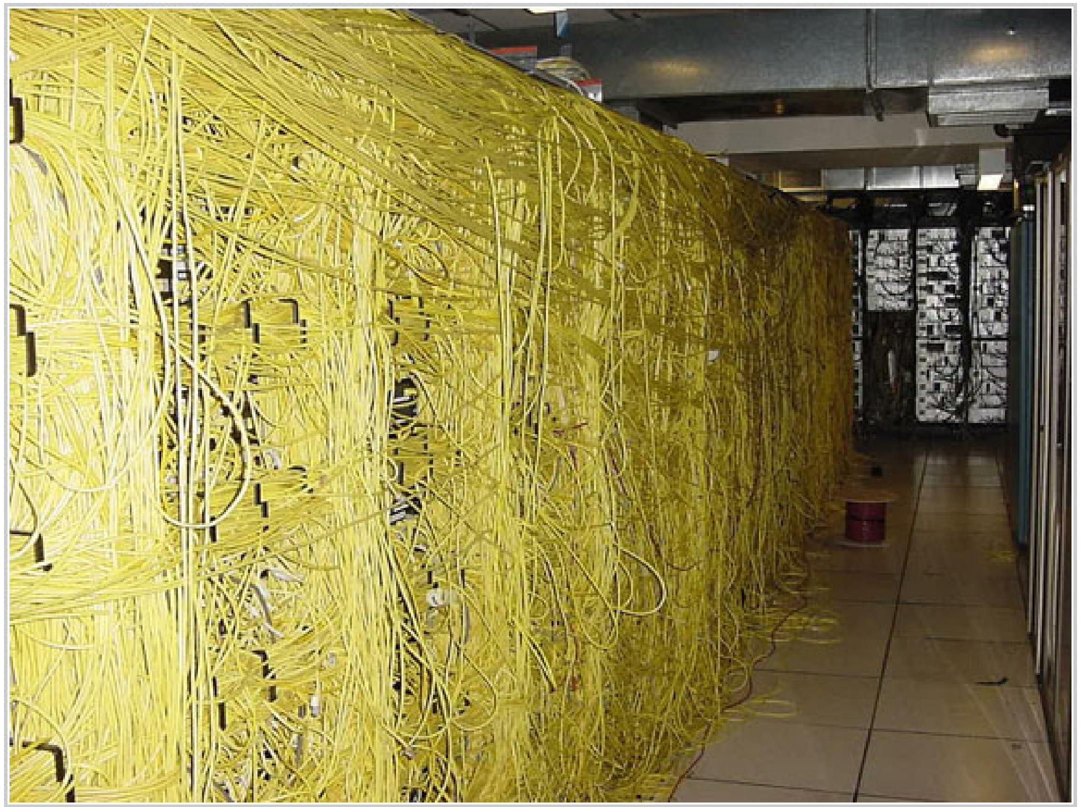
Predicting Discussions on the Social Semantic Web

Matthew Rowe, Sofia Angeletou and Harith Alani

Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom

Mass of Social Data

Social content is now published at a staggering rate....

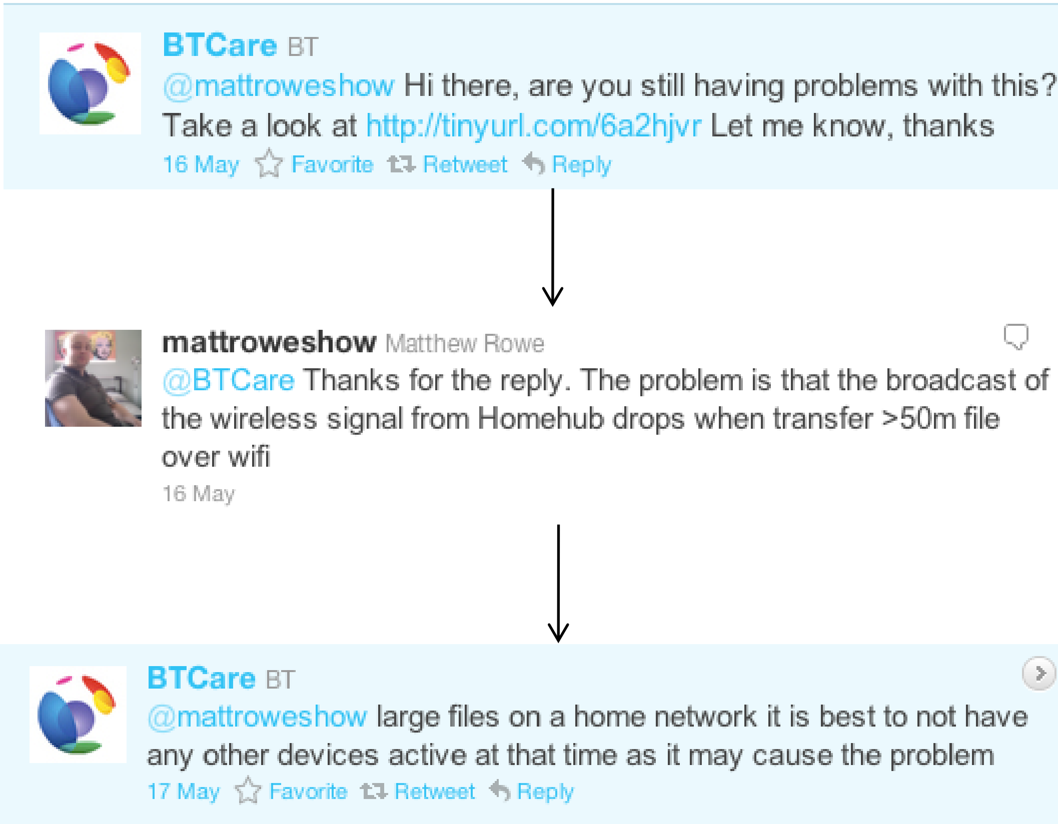


Social Data Publication Rates

- ~600 Tweets per second [1]
- ~700 Facebook status updates per second [1]
- Spinn3r dataset collected from Jan – Feb 2011 [2]
  - 133 million blog posts
  - 5.7 million forum posts
  - 231 million social media posts

[1] http://searchengineland.com/by-the-numbers-twitter-vs-facebook-vs-google-buzz-36709
[2] http://icwsm.org/data/index.php

The New Information Era



But…. Analysis is Limited

- Market Analysts
  - What are people saying about my products?
- Opinion Mining
  - How are people perceiving a given subject or topic?
- eGovernment Policy Makers
  - How is a policy or law received by the public?
  - How can I maximise feedback to my content?

Attention Economics

- Given all this data...

How do we decide on what information to focus on?

How do we know what posts will evolve into discussions?

- Attention Economics (Goldhaber, 1997)
- Need to understand **key indicators** of high-attention discussions

Discussions on Twitter

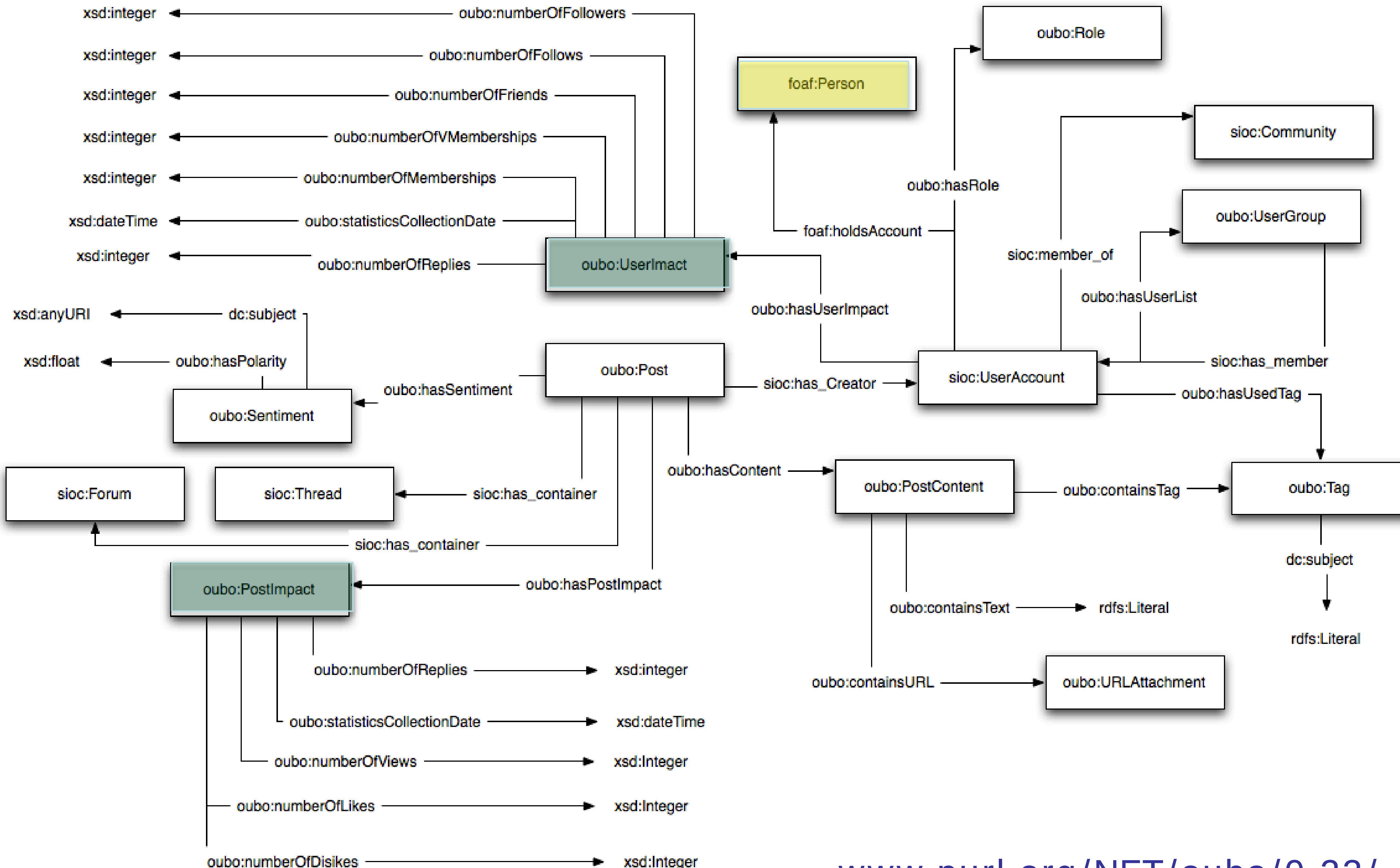
- Twitter is used as medium to:
  - Share opinions and ideas
  - Engage in discussions
    - Discussing events
    - Debating topics
- Identifying online discussions enables:
  - Up-to-date public opinion
  - Observation of topics of interest
  - Gauging the popularity of government policies
  - Fine-grained customer support

Predicting Discussions

- Pre-empt discussions on the Social Web:
  1. Identifying seed posts
     - i.e. posts that start a discussion
     - Will a given post start a discussion?
     - What are the key features of seed posts?
  2. Predicting discussion activity levels
     - What is the level of discussion that a seed post will generate?
     - What are the key factors of lengthy discussions?

The Need for Semantics

- For predictions we require statistical features
  - User features
  - Content features
- Features provided using differing schemas by different platforms
  - How to overcome heterogeneity?
- Currently, no ontologies capture such features



www.purl.org/NET/oubo/0.23/

Features

User Features		
In Degree:	Number of followers of U	#
Out Degree:	Number of users U follows	#
List Degree:	Number of lists U appears on. Lists group users by topic	#
Post Count:	Total number of posts the user has ever posted	#
User Age:	Number of minutes from user join date	#
Post Rate:	Posting frequency of the user	$\frac{PostCount}{UserAge}$
Content Features		
Post length:	Length of the post in characters	#
Complexity:	Cumulative entropy of the unique words in post p λ of total word length n and pi the frequency of each word	$\frac{\sum_{i \in [1,n]} p_i(\log \lambda - \log p_i)}{\lambda}$
Uppercase count:	Number of uppercase words	#
Readability:	Gunning fog index using average sentence length (ASL) and the percentage of complex words (PCW).	[7] $0.4(ASL + PCW)$
Verb Count:	Number of verbs	#
Noun Count:	Number of nouns	#
Adjective Count:	Number of adjectives	#
Referral Count:	Number of @user	#
Time in the day:	Normalised time in the day measured in minutes	#
Informativeness:	Terminological novelty of the post wrt other posts The cumulative tfIdf value of each term t in post p	$\sum_{t \in p} tfidf(t,p)$
Polarity:	Cumulation of polar term weights in p (using Sentiwordnet[3] lexicon) normalised by polar terms count	$\frac{Po+Ne}{\|terms\|}$

Identifying Seed Posts

- Experiments
  - Haiti and Union Address Datasets
  - Divided each dataset up using 70/20/10 split for training/validation/testing

Dataset	Users	Tweets	Seeds	Non-Seeds	Replies
Haiti	44,497	65,022	1,405	60,686	2,931
Union Address	66,300	80,272	7,228	55,169	17,875

  - Evaluated a binary classification task
    - Is this post a seed post or not?
    - Precision, Recall, F1 and Area under ROC
    - Tested: user, content, user+content features
  - Tested Perceptron, SVM, Naïve Bayes and J48

Identifying Seed Posts

(a) Haiti Dataset

		P	R	F_1	ROC
User	Perc	0.794	0.528	0.634	0.727
	SVM	0.843	0.159	0.267	0.566
	NB	0.948	0.269	0.420	0.785
	J48	0.906	0.679	0.776	0.822
Content	Perc	0.875	0.077	0.142	0.606
	SVM	0.552	0.727	0.627	0.589
	NB	0.721	0.638	0.677	0.769
	J48	0.685	0.705	0.695	0.711
All	Perc	0.794	0.528	0.634	0.726
	SVM	0.483	0.996	0.651	0.502
	NB	0.962	0.280	0.434	0.852
	J48	0.824	0.775	0.798	0.836

(b) Union Address Dataset

		P	R	F_1	ROC
User	Perc	0.658	0.697	0.677	0.673
	SVM	0.510	0.946	0.663	0.512
	NB	0.844	0.086	0.157	0.707
	J48	0.851	0.722	0.782	0.830
Content	Perc	0.467	0.698	0.560	0.457
	SVM	0.650	0.589	0.618	0.638
	NB	0.762	0.212	0.332	0.649
	J48	0.740	0.533	0.619	0.736
All	Perc	0.630	0.762	0.690	0.672
	SVM	0.499	0.990	0.664	0.506
	NB	0.874	0.212	0.341	0.737
	J48	0.890	0.810	0.848	0.877

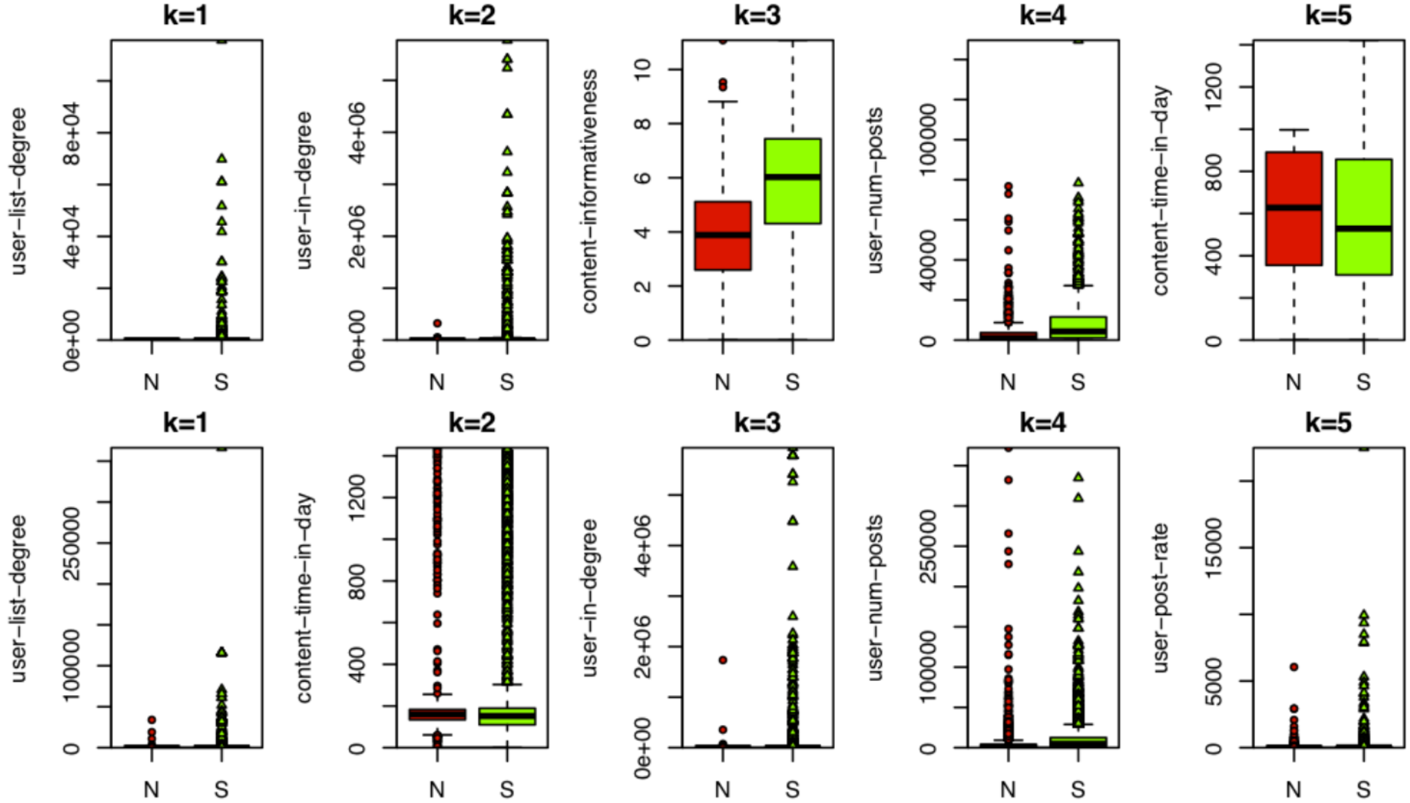
Identifying Seed Posts

- What are the most important features?

Rank	Haiti	Union Address
1	user-list-degree (0.275)	user-list-degree (0.319)
2	user-in-degree (0.221)	content-time-in-day (0.152)
3	content-informativeness (0.154)	user-in-degree (0.133)
4	user-num-posts (0.111)	user-num-posts (0.104)
5	content-time-in-day (0.089)	user-post-rate (0.075)
6	user-post-rate (0.075)	user-out-degree (0.056)
7	content-polarity (0.064)	content-referral-count (0.030)
8	user-out-degree (0.040)	user-age (0.015)
9	content-referral-count (0.038)	content-polarity (0.015)
10	content-length (0.020)	content-length (0.010)
11	content-readability (0.018)	content-complexity (0.004)
12	user-age (0.015)	content-noun-count (0.002)
13	content-uppercase-count (0.012)	content-readability (0.001)
14	content-noun-count (0.010)	content-verb-count (0.001)
15	content-adj-count (0.005)	content-adj-count (0.0)
16	content-complexity (0.0)	content-informativeness (0.0)
17	content-verb-count (0.0)	content-uppercase-count (0.0)

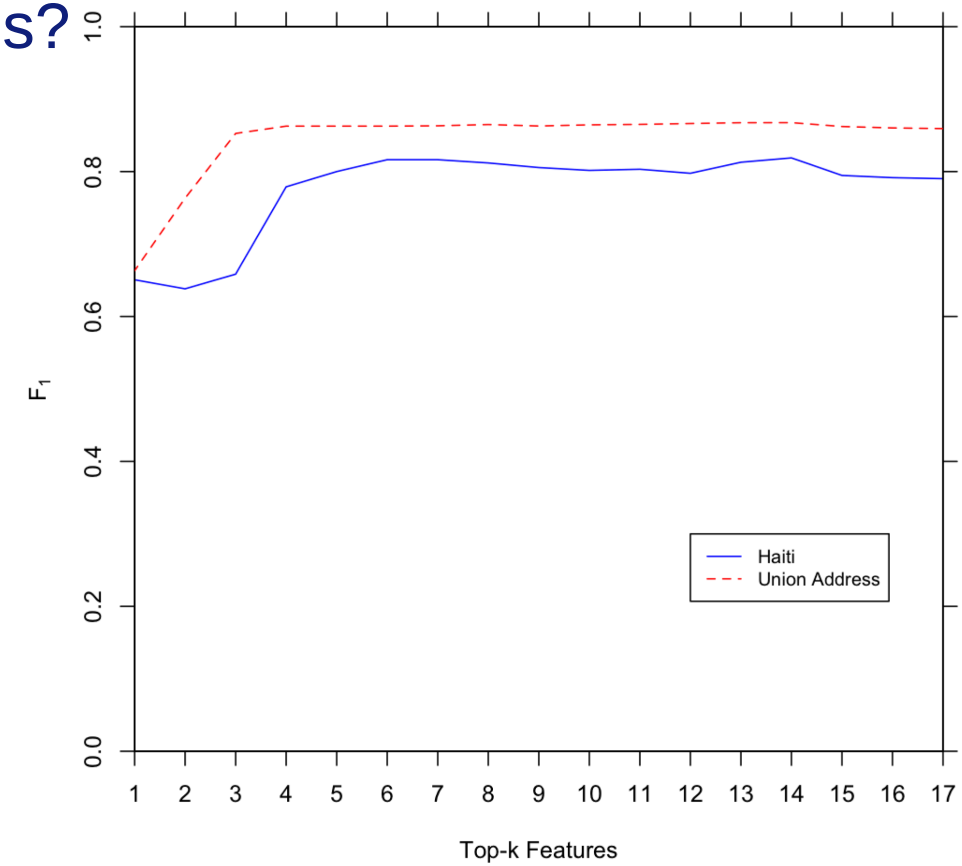
Identifying Seed Posts

- What is the correlation between seed posts and features?



Identifying Seed Posts

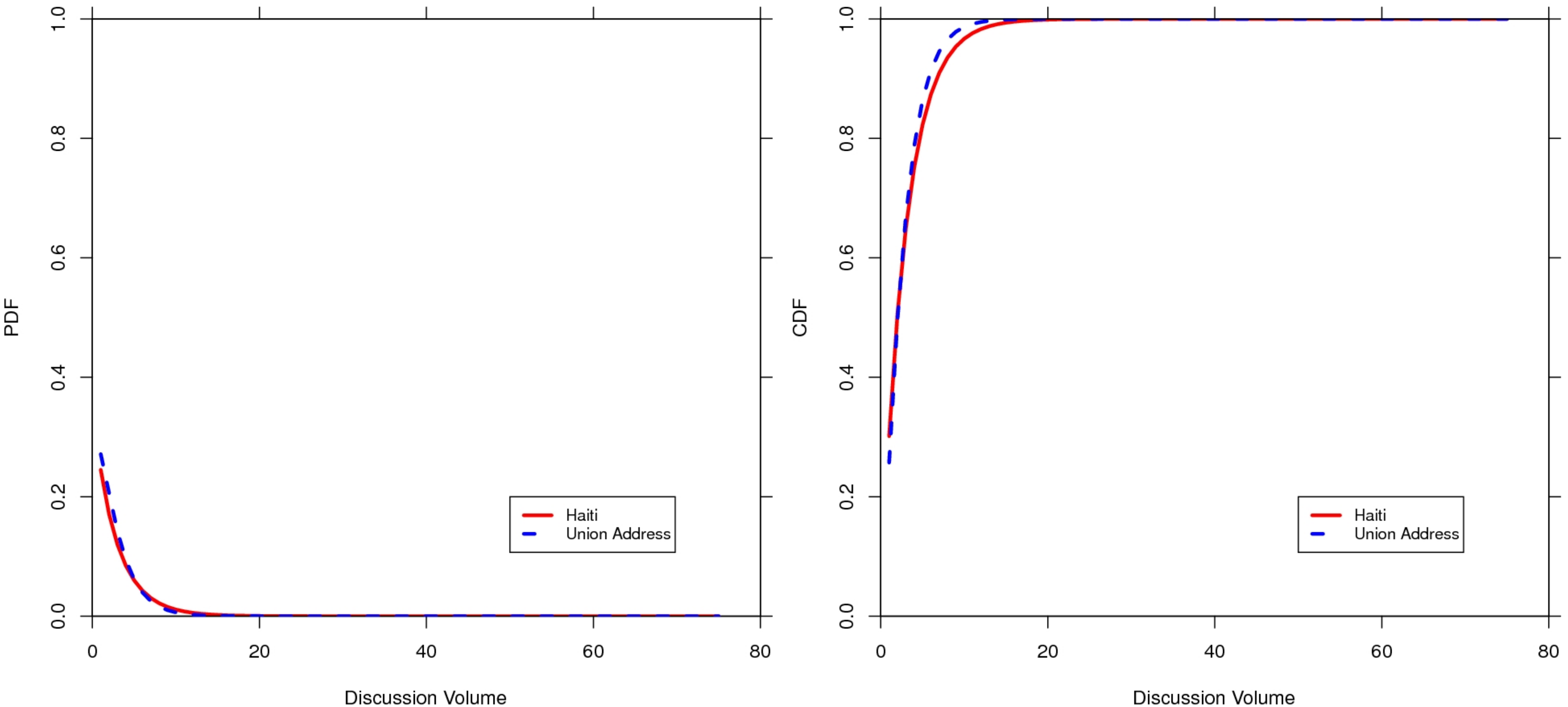
- Can we identify seed posts using the top-k features?



Predicting Discussion Activity

- From identified seed posts:
  - Can we predict the level of discussion activity?
  - How much activity will a post generate?
- [Wang & Groth, 2010] learns a regression model, and reports on coefficients
  - Identifying relationship between features
- We do something different:
  - *Predict the volume of the discussion*

Predicting Discussion Activity

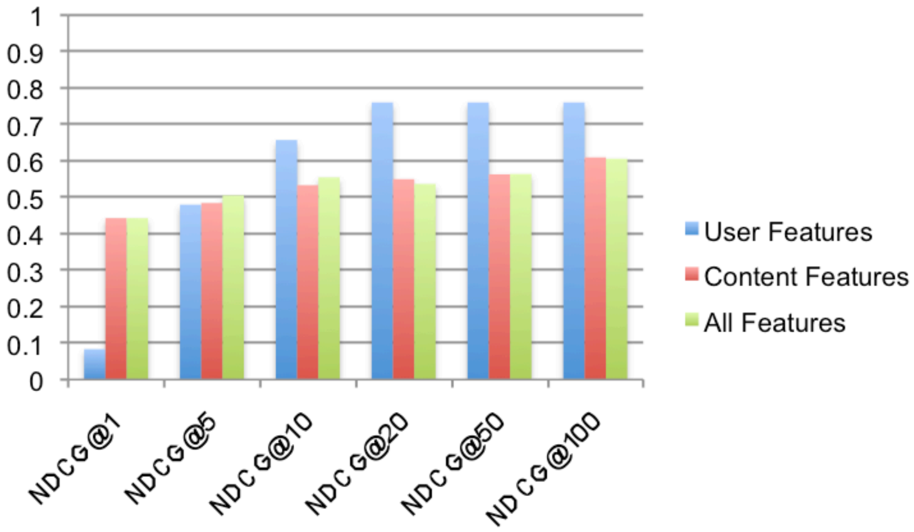


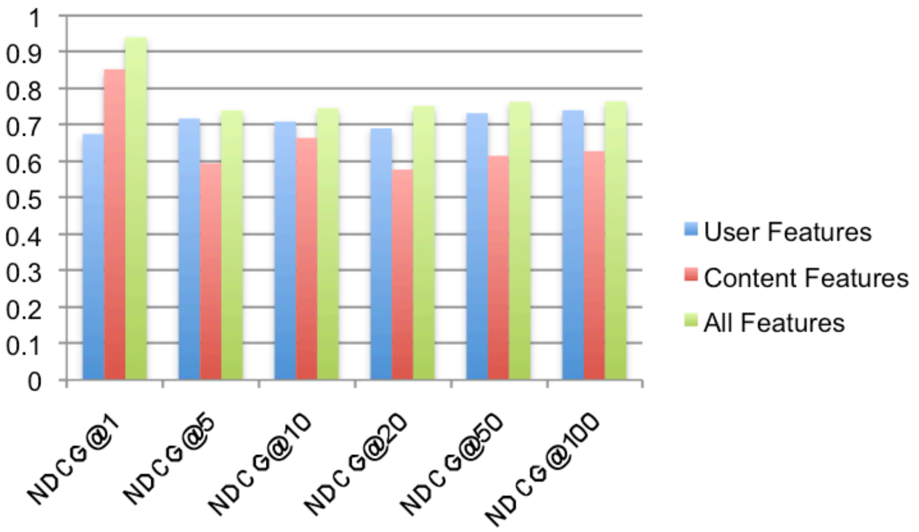
Predicting Discussion Activity

- Compare rankings
  - Ground truth vs predicted
- Experiments
  - Using Haiti and Union Address datasets
  - Evaluation measure: Normalised Discounted Cumulative Gain
    - Assessing nDCG@k where k={1,5,10,20,50,100)
  - Tested Support Vector Regression with:
    - user, content, user+content features

Dataset	Train Size	Test Size	Test Vol Mean	Test Vol SD
Haiti	980	210	1.664	3.017
Union Address	5,067	1,161	1.761	2.342

Predicting Discussion Activity





	user-num-posts	user-out-degree	user-in-degree	user-list-degree	user-age	user-post-rate
Haiti	-0.0019	+ 0.001	+ 0.0016	+ 0.0046	+ 0.0001	+ 0.0001
Union	-0.0025	+ 0.0114	+ 0.0025	+ 0.0154	-0.0003	-0.0002

Findings

- User reputation and standing is crucial
  - eliciting a response
  - starting a discussion
- Greater broadcast capability = greater likelihood of response
  - More listeners = more discussion
- Activity levels influenced by out-degree
  - Allow the poster to see response from ‘respected’ peers

Conclusions

- Pre-empt discussions to empower
  - Market analysts
  - Opinion mining
  - eGovernment policy makers
- Behaviour ontology
  - Captures impact across platforms
- Approach accurately predicts:
  - Which posts will yield a reply, and;
  - The level of discussion activity

Current and Future Work

- Experiments over a forum dataset
  - Content features >> user features
  - Different platform dynamics
- Extend experiments to a random Twitter dataset
- Extension to behaviour ontology
  - Captures concentration
  - i.e. focus of a user on specific topics
- Categorising users by role
  - Based on observed behaviour

Questions?

people.kmi.open.ac.uk/rowe
m.c.rowe@open.ac.uk
@mattroweshow