

A faint, light gray network graph is visible in the background, consisting of numerous small square nodes connected by thin lines, forming a complex web-like structure.

Information That Matters: Investigating Relevance of Entities in Social Media Networks

Munmun De Choudhury

PhD Candidate, Computer Science

Arizona State University

<http://www.public.asu.edu/~mdechoud/>

What do I do?

Modern Social Interactional Modes



Some of my most fav pics are here!

A close-up, slightly blurred photograph of a person's hands holding a smartphone. The phone's screen displays a list of text, possibly a contact list or a document. The background is out of focus, showing warm, bokeh-like light spots. The overall tone is warm and intimate, suggesting a personal or social context.

**Understanding the
dynamics and
impact of our online
social interactions**

And because...

140 characters
can cause
revolutions

During the elections in Iran



And during the earthquake in Haiti



- 1) Sustainability of culture
in the digital society
- 2) Next-generation
interactive social
information systems

And why
should *you*
care?

Bing, Windows Live Mail



Viral Marketing, Advertizing Campaigns

Xbox Live, Microsoft Office Suite, Visual Studio



Collaboration in Organizations

Bing, Windows phone



Better Interface Design

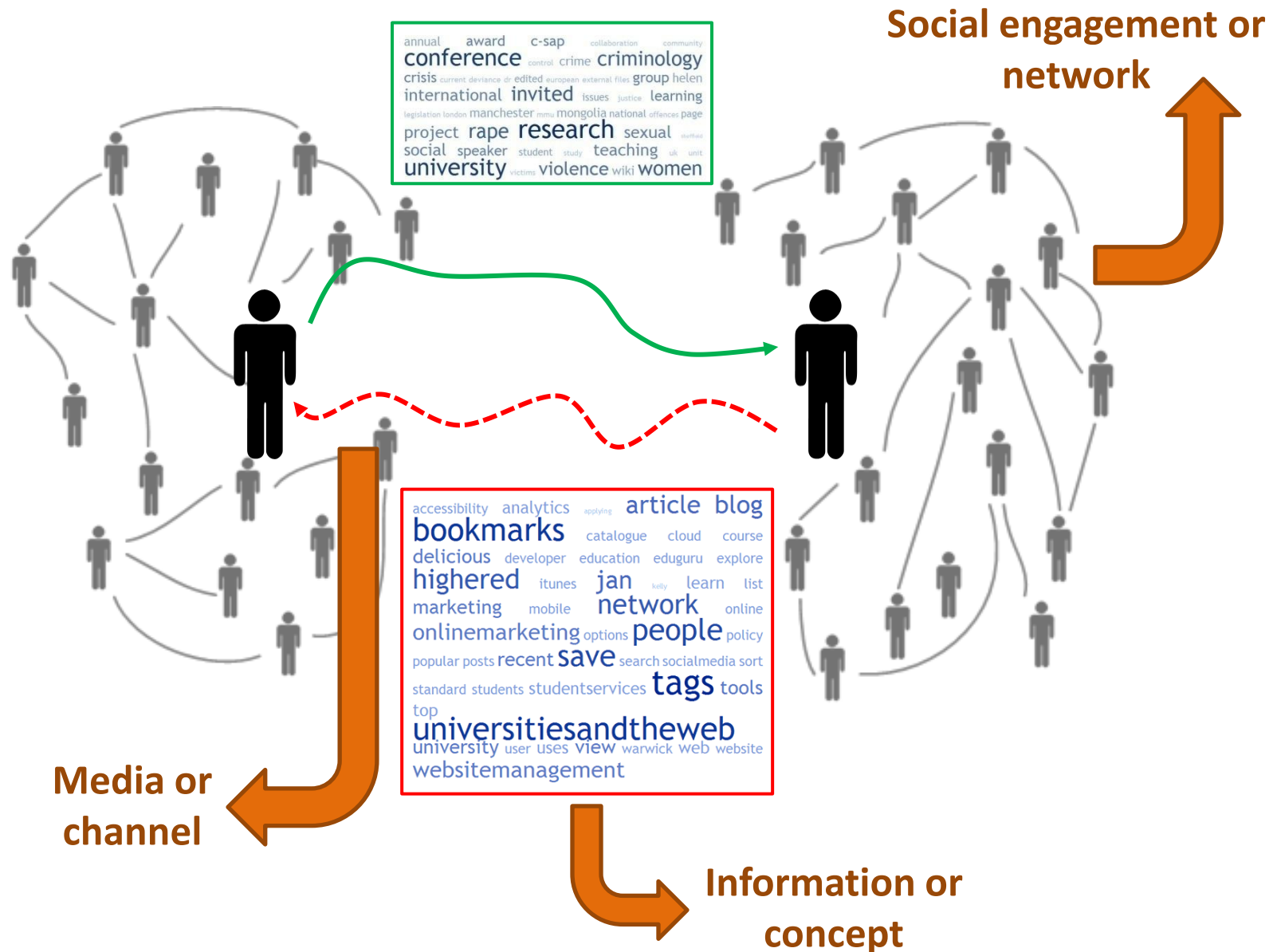
**Bing, Windows Live Mail,
Windows phone, xbox Live**

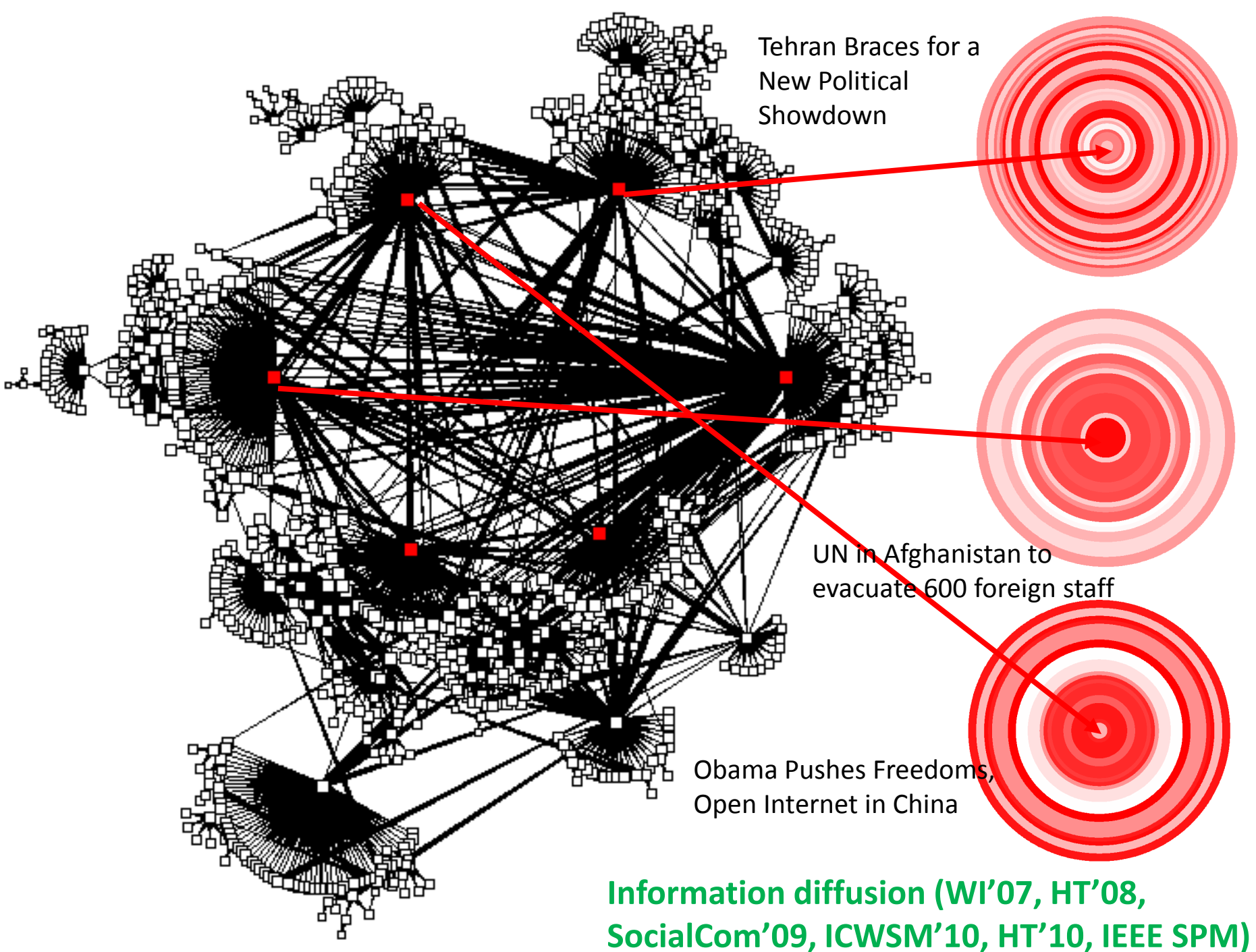


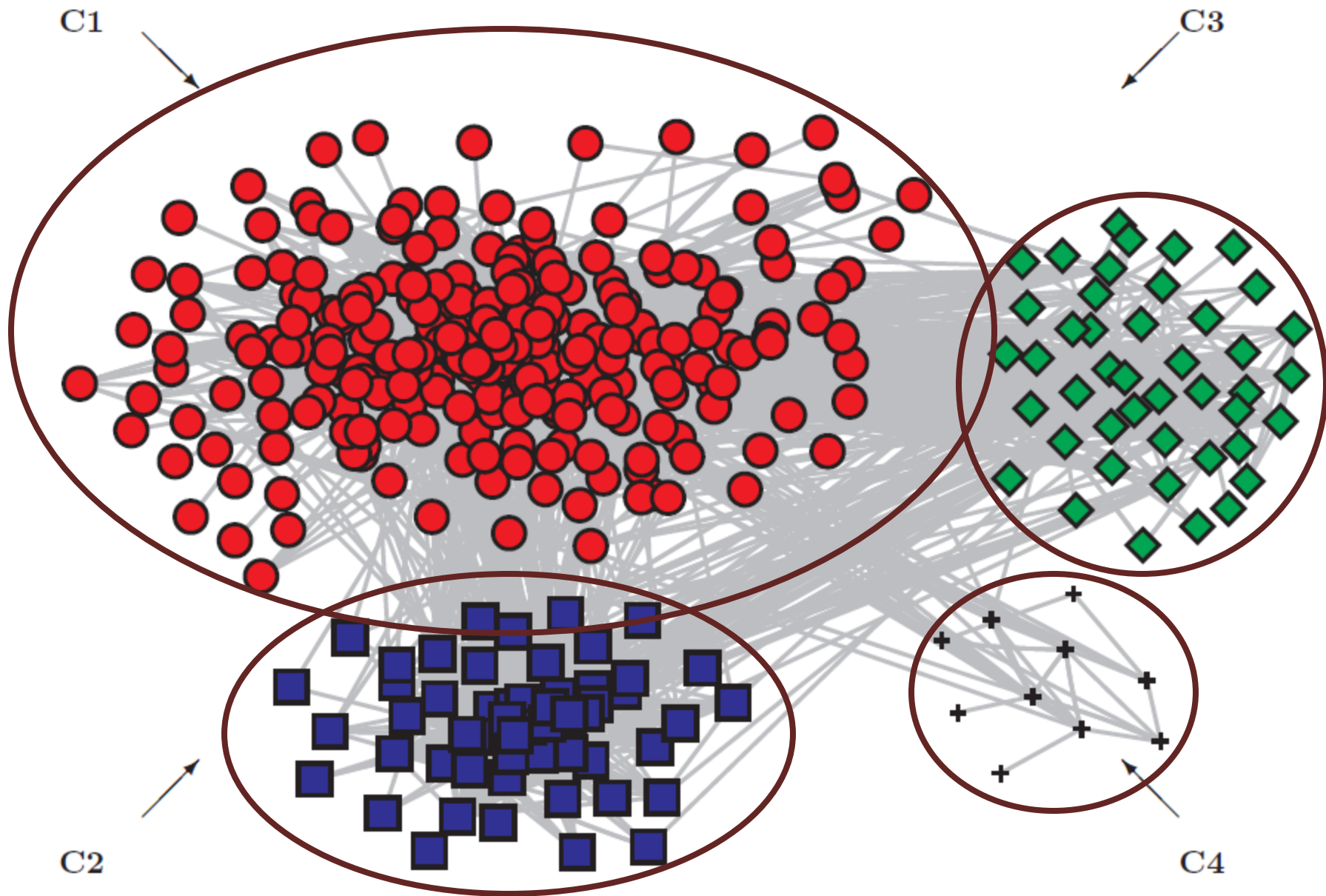
Distributed Social Search

But how do we model
and analyze our
interactions to address
these applications?

Alice and Bob are two users of the xbox Live gaming software







Community evolution (HT'08, CIKM'08, HT'09, ICME'09, ACM TOIS)



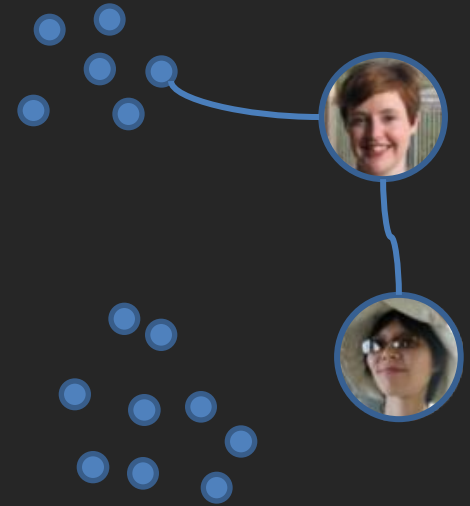
**However the social
web is changing at a
fast rate**

And *what* exactly
is changing?

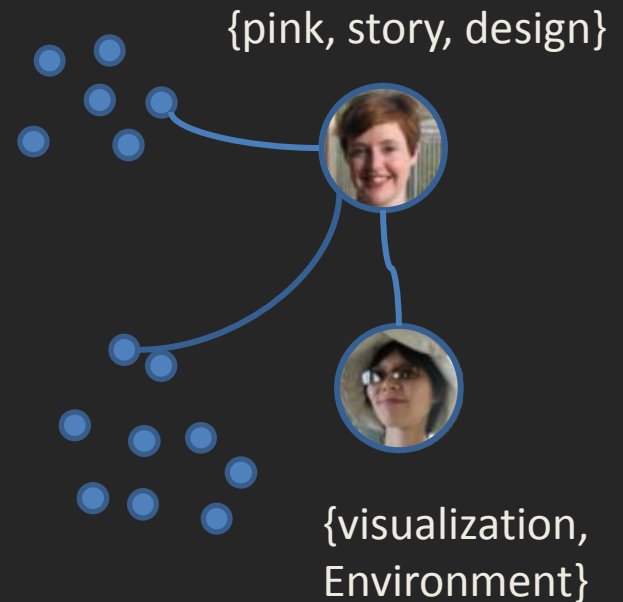
New people
appear



New ties are formed

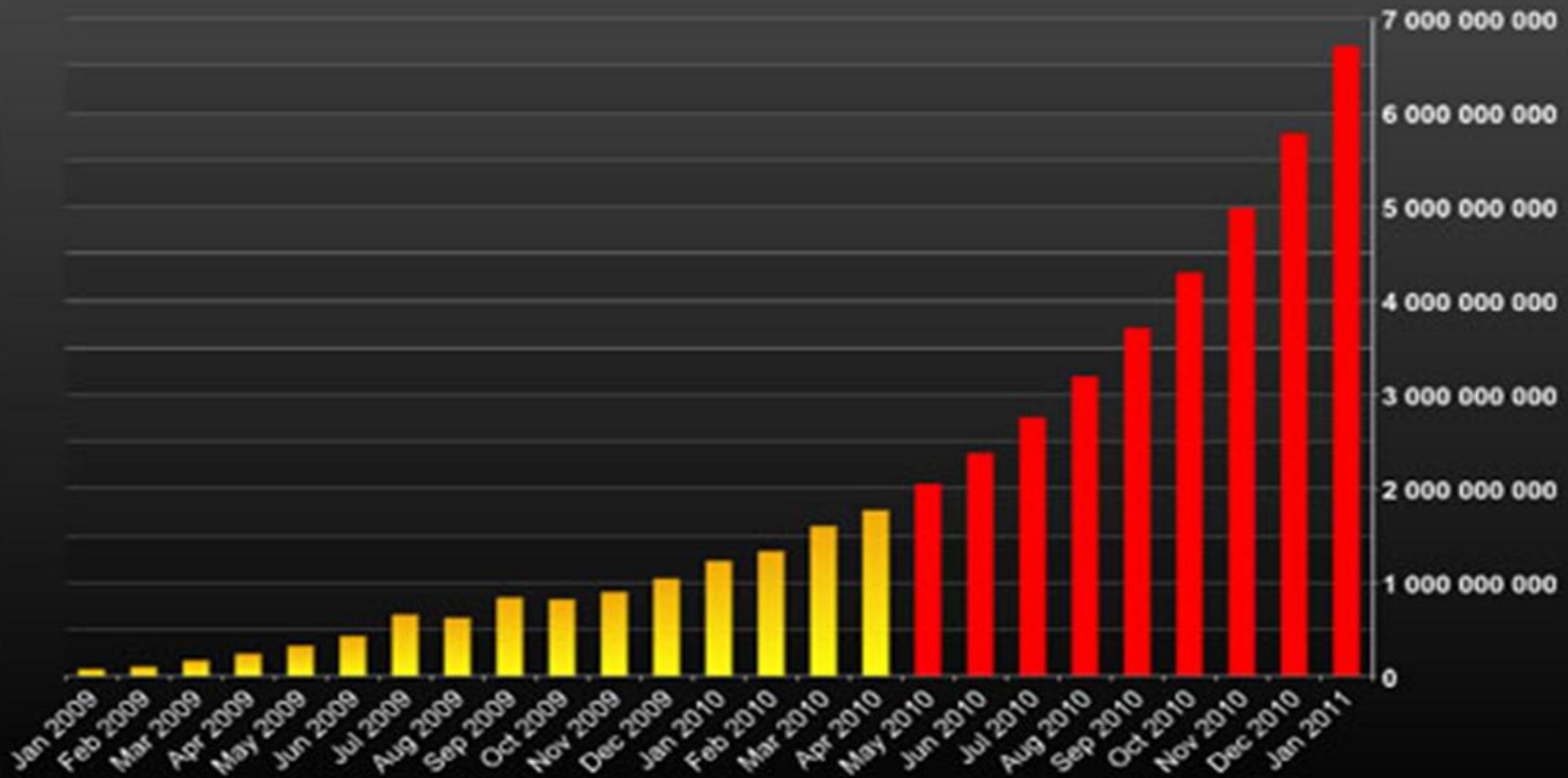


New
interactional
data appears
too!



By April 2010, <http://www.twitter.com/> was receiving over 600 million search queries per day (Huffington Post).

Tweets per month on Twitter: Predicted growth



We are attracted
to social media,
in part due to
large scale
datasets



A young girl with curly hair is looking upwards with a wide-eyed, open-mouthed expression of awe or surprise. The background is bright and out of focus, showing green foliage and a hand pointing towards her. Overlaid on the image is the text "Is there something more fundamental happening here than just scale?" in a bold, red, sans-serif font.

**Is there something
more fundamental
happening here than
just scale?**

A person with glasses is shown in profile, looking at a laptop screen. The screen displays lines of code. The person's hand is near their face, possibly resting their chin. The background is dark. Overlaid on the image is large white text, with the word 'matters' in yellow italics.

This talk is
about sampling
for information
that *matters*

Two simple questions





How do we infer
meaningful human
networks?

([WWW'10](#)) – at Yahoo! Research

How do we
identify
valuable social
media content?



([WWW'10](#), [HT'10](#), in prep.) – at Yahoo! Research, Microsoft Research

Question I

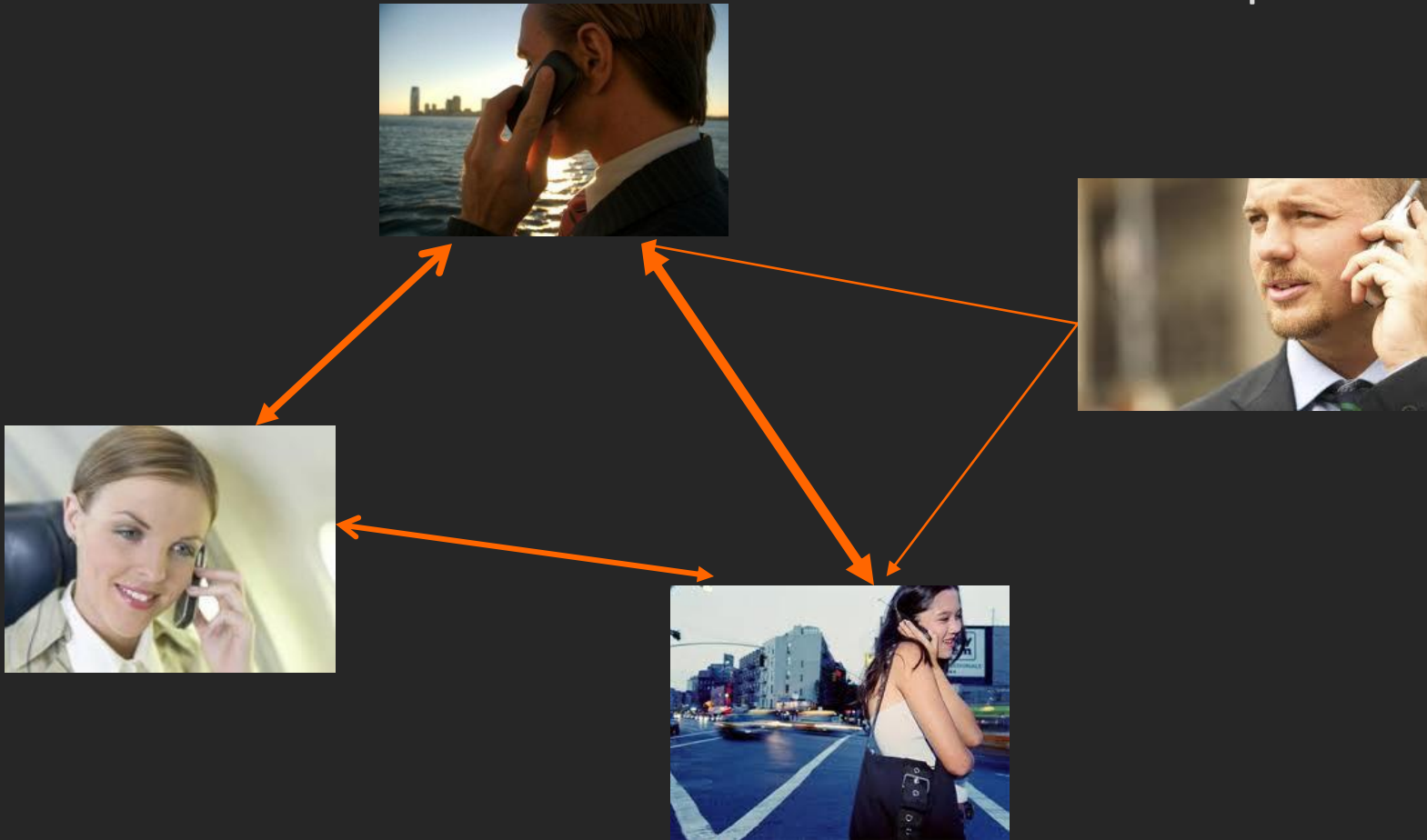
- With Winter Mason, Jake Hofman and Duncan Watts, during internship at Yahoo! Research, summer 2009

How to choose a
relevant tie?

Social ties from communication data

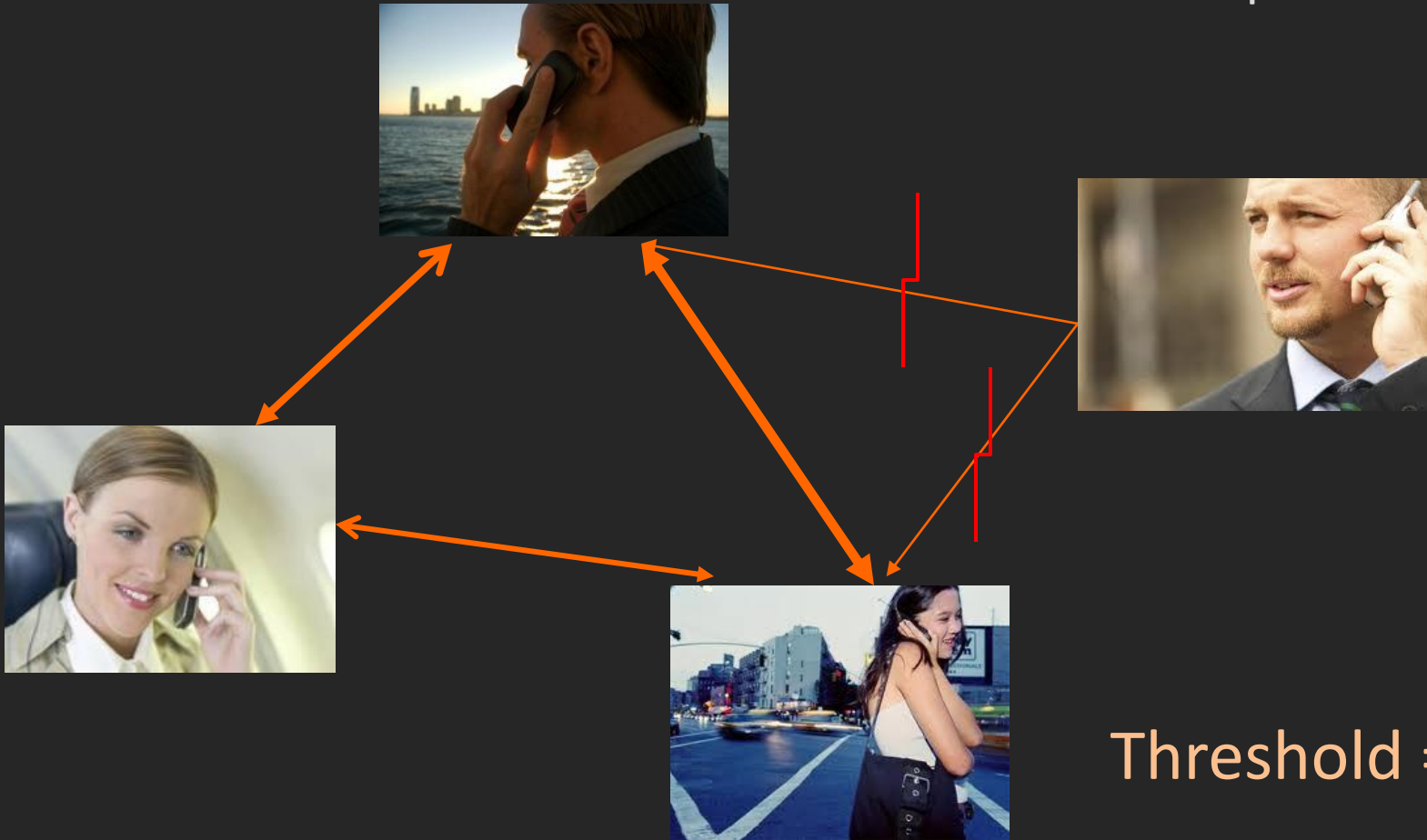
- Reasonable definitions:
 - At least one communication in past year
 - Average of one communication every week
 - One reciprocated communication in past month
- What is the research question?
 - Search on network
 - Information diffusion
 - Uncovering hidden node properties
- *Our method to find relevant ties: define a minimum threshold*

Four Windows phone users



Defining a minimum threshold

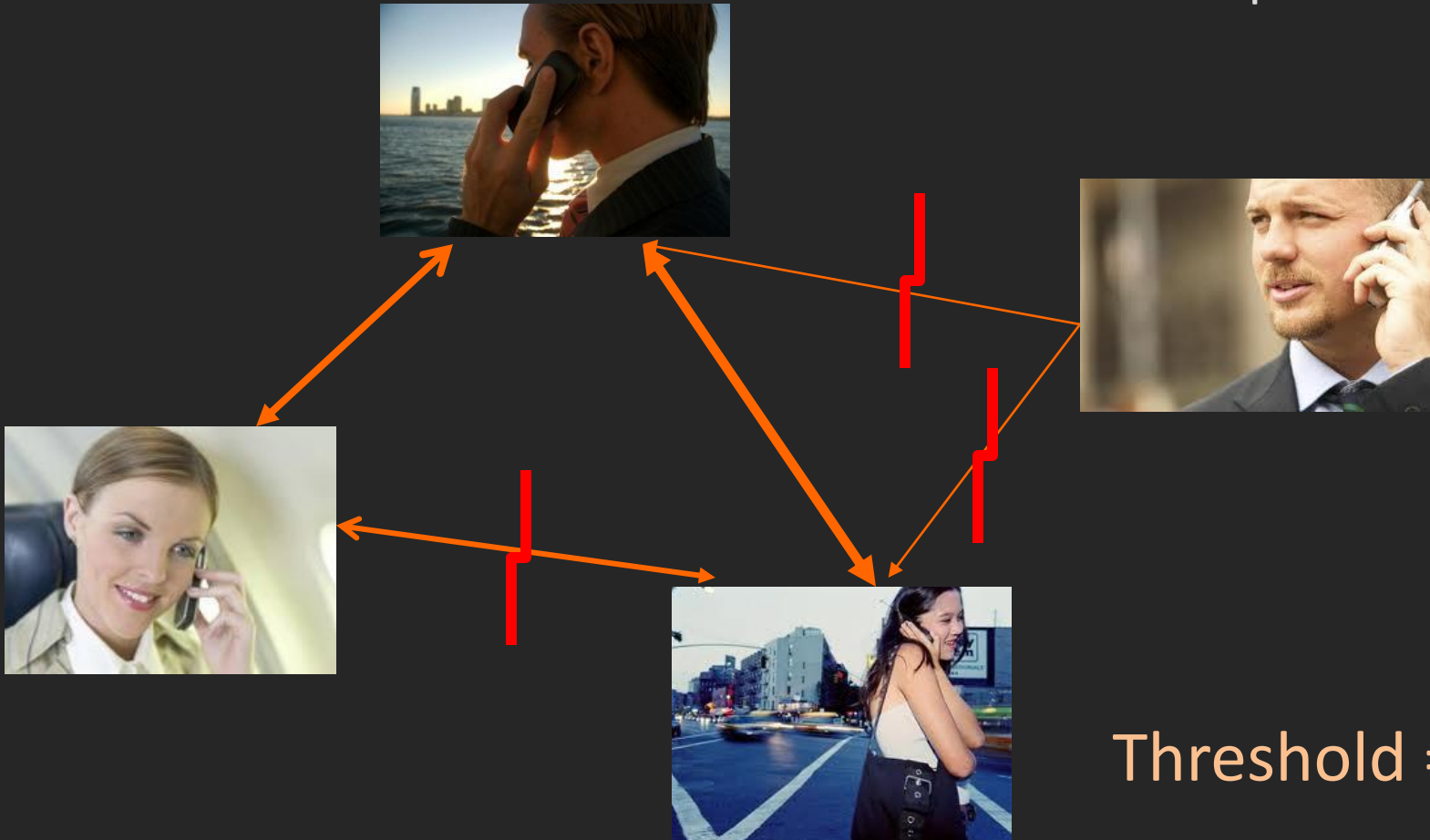
Four Windows phone users



Threshold =

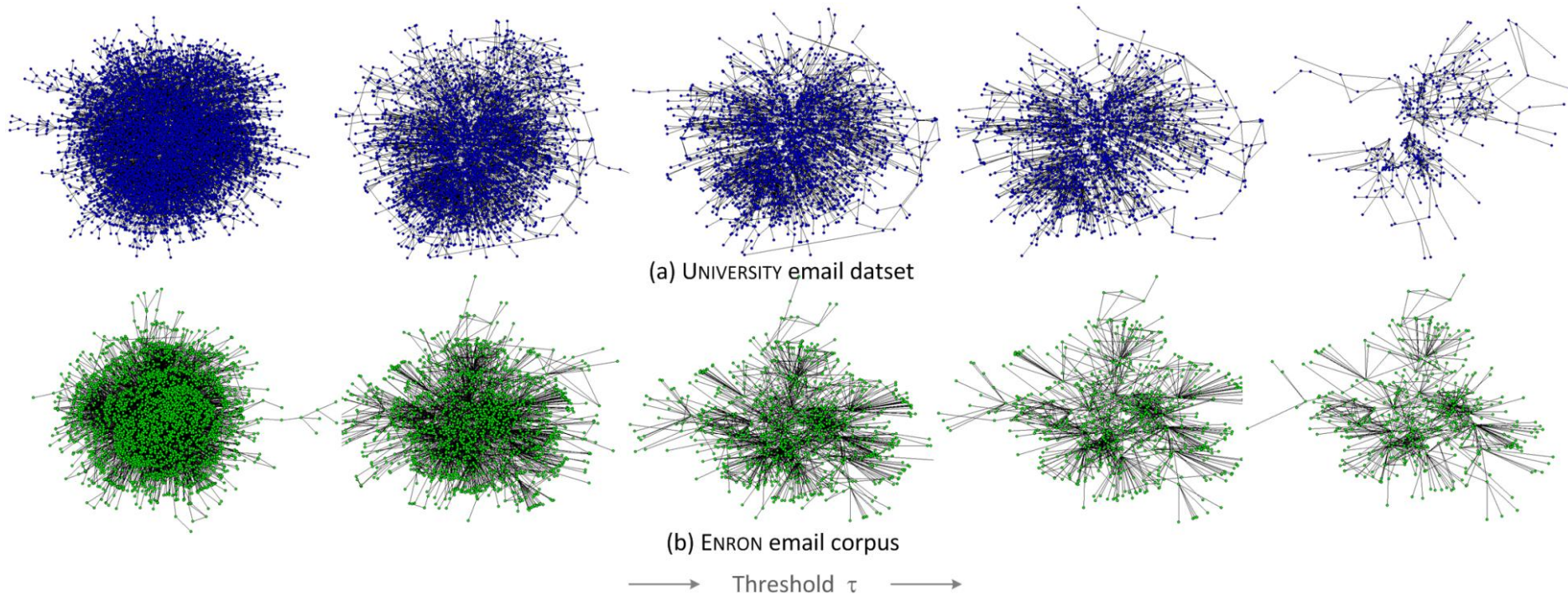
Defining a minimum threshold

Four Windows phone users



Threshold =

Defining a minimum threshold



Why inferring the relevant tie matters

Our Contributions

- **Goal:**
 - Infer networks for various definitions of “threshold” over a tie
 - Study the impact of different thresholded networks on:
 - descriptive statistics and
 - ability of the network in predicting node characteristics
- **What insights can we gain on what are the optimal thresholds to define relevant ties?**

Datasets

- **University Email**

- a complied registry of all email (incoming and outgoing, as recorded in server logs) associated with individuals at a large university in the US, comprising undergraduate and graduate students, faculty, and staff
- Focus on a consistent user set across all semesters - 19,817 individuals
- 1.09M emails; disregard emails involving non-university domain
- 2 years (6 semesters – in the order Fall, Spring, Summer)
- PS: content of emails not available

- **Enron Email**

- a repository of the emails exchanged internally among the employees at the Enron Corporation, obtained through a subpoena as part of an investigation by the Federal Energy Regulatory Commission (FERC) and then made public
- 4,736 individuals
- 1.06M emails
- 4 years (1998-2002)

“Thresholded” Networks

- **Edge definition:**

- Symmetric edge based on the frequency of email communication
- Geometric mean of the annualized rate of messages exchanged over the span of two and four years respectively. For users u_i and u_j :

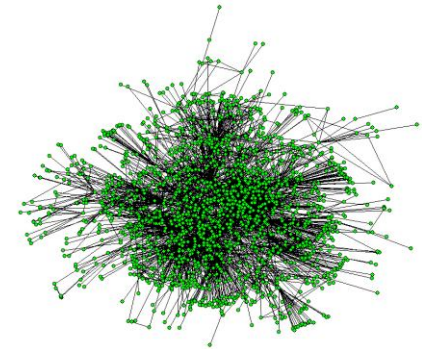
$$e_{ij} = \sqrt{w_{ij}w_{ji}}$$

- **Edge threshold:**

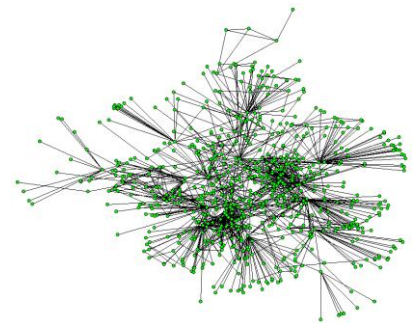
- Minimum of τ emails between each pair of individuals, over a period of time T . Hence we construct the social graph $G(V,E;\tau)$ such that,

$$e_{ij} \in E \text{ if and only if, } e_{ij} \geq \frac{\tau}{T}.$$

- Family of networks: $\{G(\tau_1), G(\tau_2), \dots, G(\tau_K)\}$



$\tau=5$ emails per year

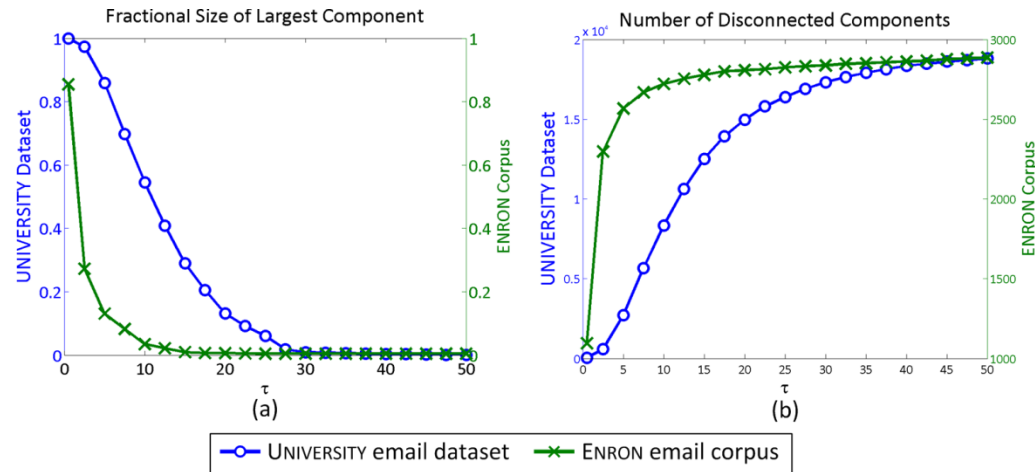


$\tau=15$ emails per year

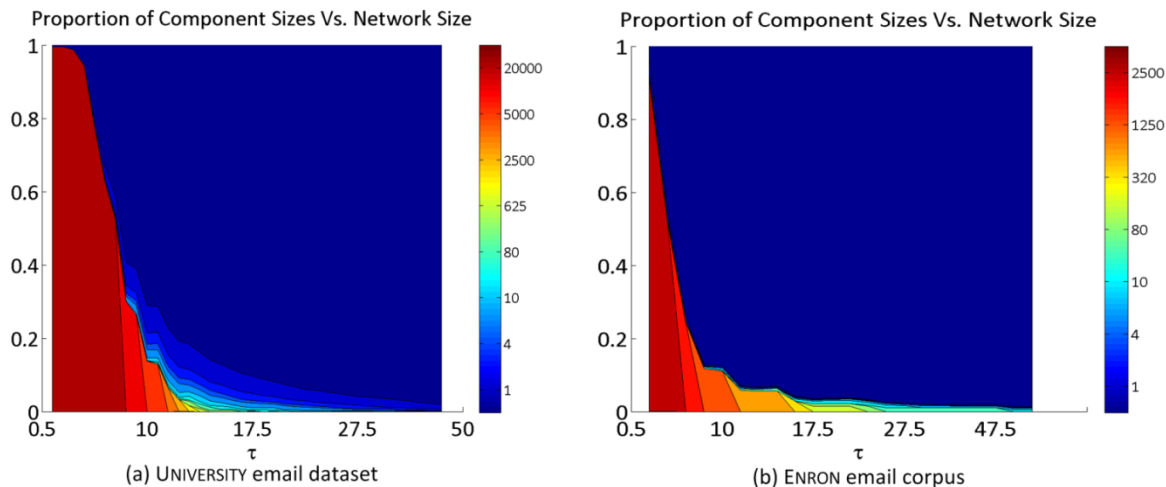
Network Descriptive Statistics

Global Network Features

- Number of connected components

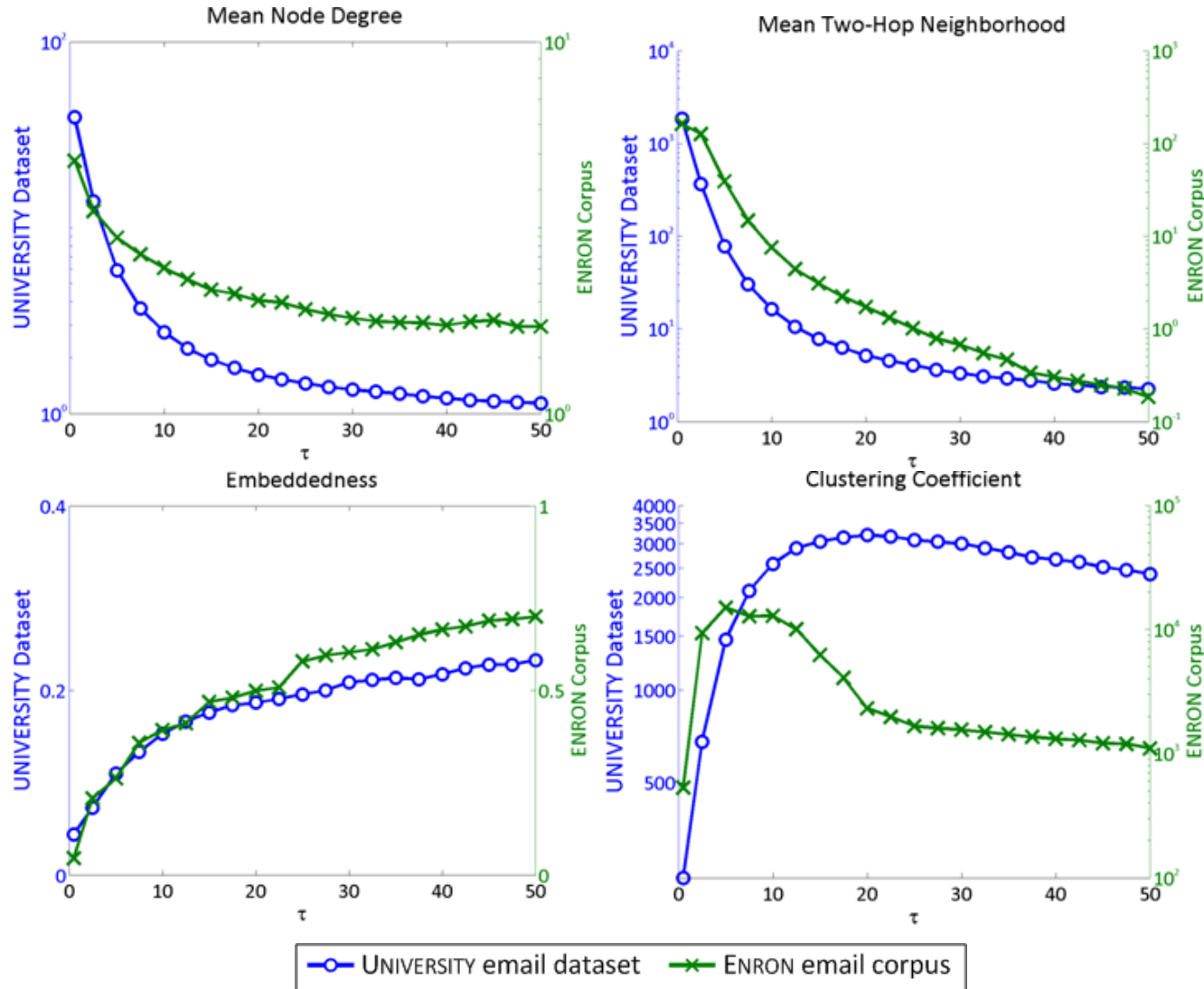


- Relative sizes of components



Local Network Features

- Reach, closure, bridge measures



How to choose the
right threshold?

Premise of Prediction

Define an edge according to the research problem of interest

- Making predictions on the inferred networks based on the structural properties of the network:
 - Normalized clustering coefficient
 - Node degree
 - Embeddedness
 - Two-hop neighborhood, etc.
- Which network (defined by a certain threshold) yields the best prediction?

Prediction Tasks: Node Status/Gender

- Given feature set of structural features & mean edge weight of neighbors with attribute i :

where ω_j gives the mean edge weight of u_i with respect to the neighbors having attribute value j ($1 \leq j \leq q$) and $N_i(a_j)$ is the subset of i 's neighbors whose attribute value is j

- Also consider an unweighted version with all $\omega_j=1$
- Split into training (90%) and test (10%) sets
- Use SVM (Support vector machine based attribute prediction) with Gaussian RBF kernel, learn parameters & kernel width with k -fold cross-validation ($k=10$ in this work)

Prediction Tasks: Future Communication

- To predict activity of a user u_i at time t_{m+1} , we use a similar feature-based representation of u_i in the network $G(\tau)$, i.e.
 - the structural features
 - the mean weighted activities of her neighbors from time t_0 to t_m
 - we augment the feature space by using u_i 's communication from t_0 to t_m
- We fit a linear model of communication activity as a function of the node level features $\mathbf{F}_{0:m}^\tau$:

$$A_m = \beta_{0:m}^\tau \cdot \mathbf{F}_{0:m}^\tau + \varepsilon_{0:m}^\tau, \text{ where } \varepsilon_{0:m}^\tau \text{ is additive noise.}$$

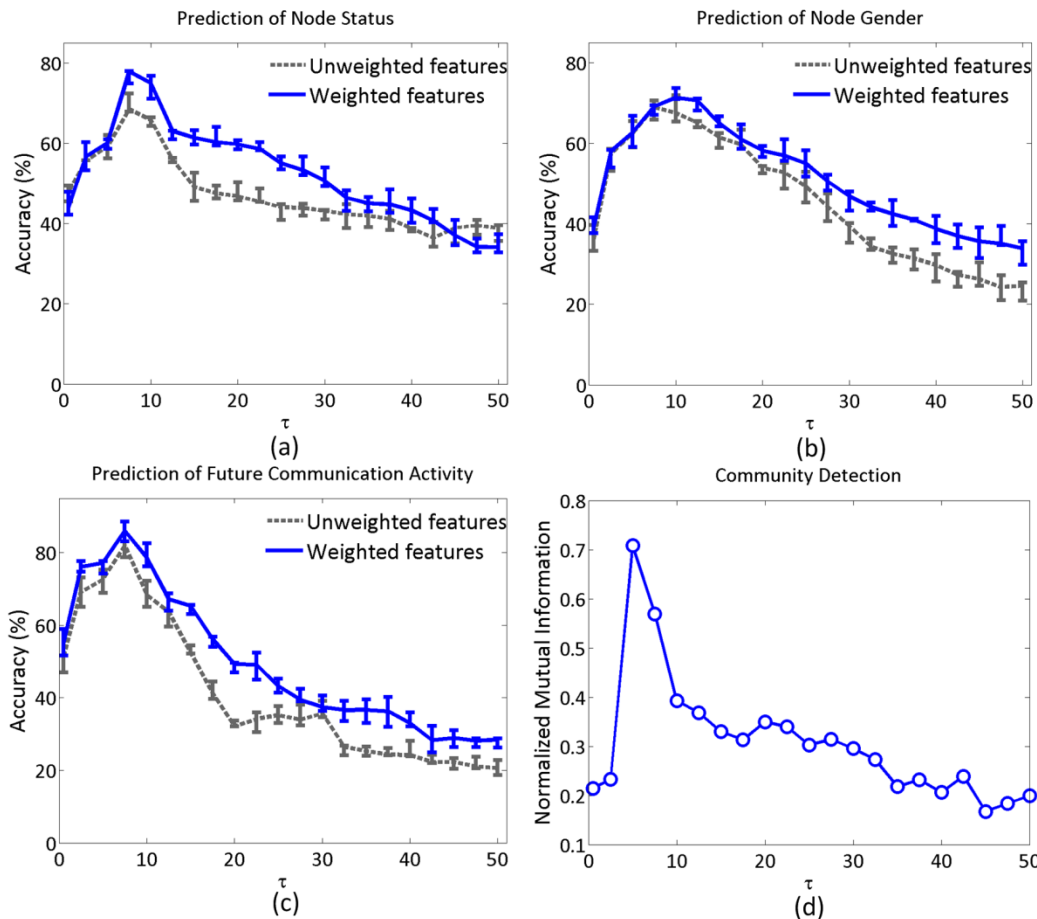
- The best-fit coefficients $\beta_{0:m}^\tau$ are used along with the feature vector at t_{m+1} , to predict future node activity given as $A'_{m+1} \in \mathbf{R}^{1 \times |V|}$

Prediction Tasks: Community Detection

- Fit a stochastic block model to $G(\tau)$ using variational Bayes inference [Hofman et al. 2008]
- **Method:**
 - Assume each node u_i belongs to one of the Z latent groups/“blocks” (or school assignments), given as z_i with probability π_μ , $\mu=1,2,\dots,Z$
 - If the nodes u_i and u_j are in the same group ($z_i=z_j$), an edge exists between them with probability ϑ_+ ; if they are in different groups ($z_i \neq z_j$), an edge exists between them with probability ϑ_-
 - Given only the observed edges $e_{ij} \in E_s$ in the graph $G(\tau)$, distributions over the group assignments $p(z_i)$ are inferred via variational Bayesian inference
- Compare soft assignments to actual school affiliation using normalized mutual information
- In our experiments, $Z=5$ for the University dataset

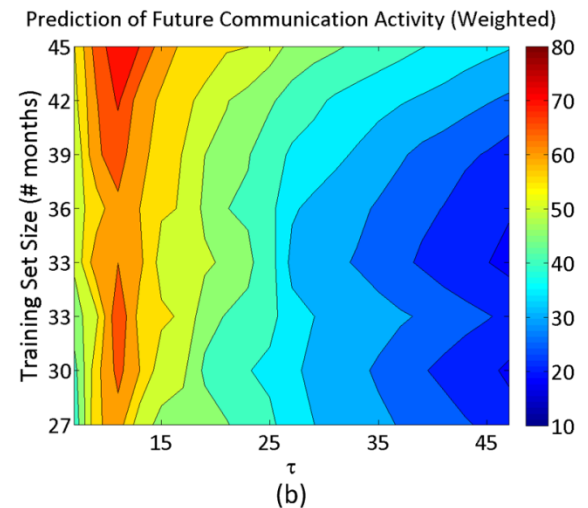
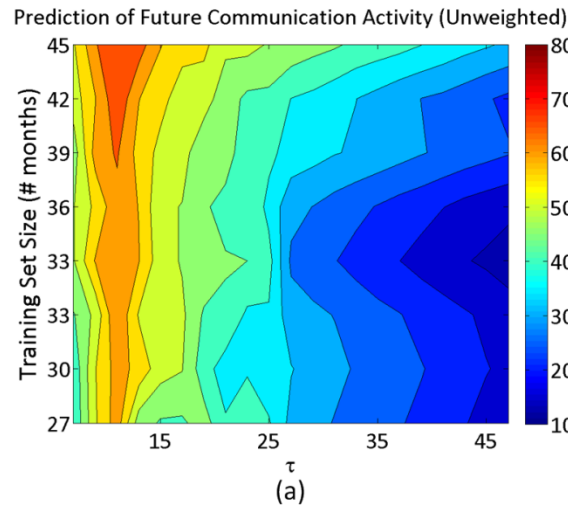
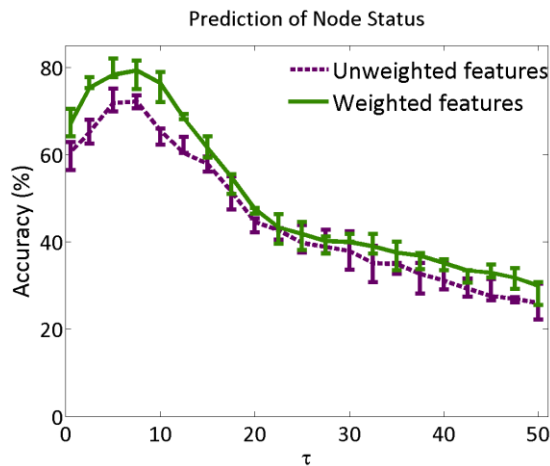
Experimental Results (University Email)

- Peak accuracy in different prediction tasks occurs at a non-trivial τ .
- There is ~30-40% boost in accuracy over unthresholded network.



Experimental Results (Enron Email)

- Best accuracy occurs at $\tau=7.5$ for the two prediction tasks
- Accuracy increases from $\sim 60\%$ to $\sim 70\%$ from unthresholded graph to optimal τ for unweighted features, and $\sim 65\%$ to $\sim 80\%$ for weighted features

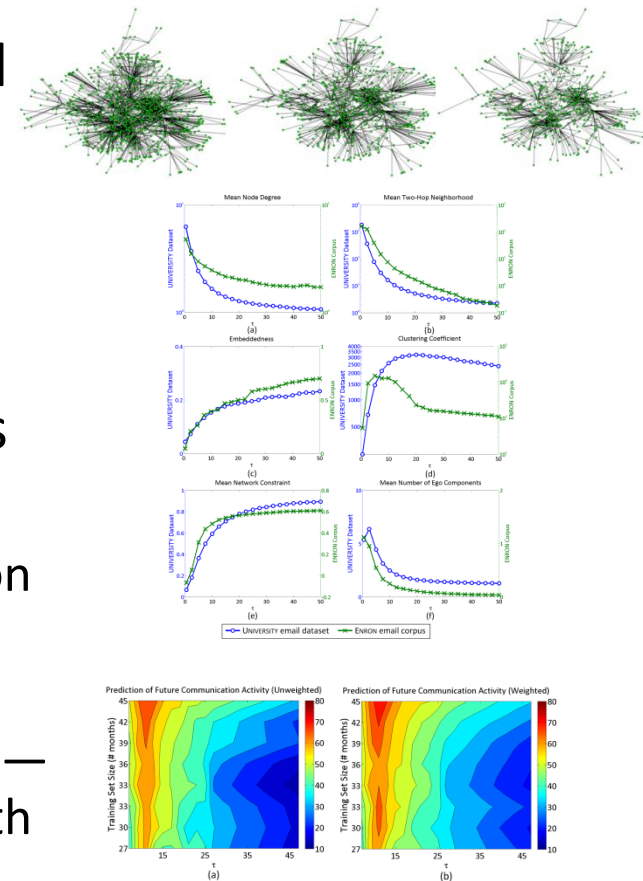


Observations

- Finding Optimal Threshold
 - Accuracy maximized at non-obvious point
 - Increase in accuracy from unthresholded graph as much as ~**30%**
 - Increase in accuracy exists even including information about weights at edges; therefore **deleting edges removes noise** (increasing signal)
- Optimal threshold at consistent value
 - For different prediction tasks
 - For different data sets

Conclusions

- Network analysis of communication data takes as input some set of observations and infers from these data a set of relations to which social and psychological meaning is attached
 - Network inference procedure largely ad-hoc
- We have addressed a narrow version of this general problem:
 - how to determine an optimal threshold condition for edges so as to predict particular node attributes (e.g. gender, status) or behavior
 - The prediction accuracies peak in a non-obvious—yet relatively narrow, threshold range across both datasets



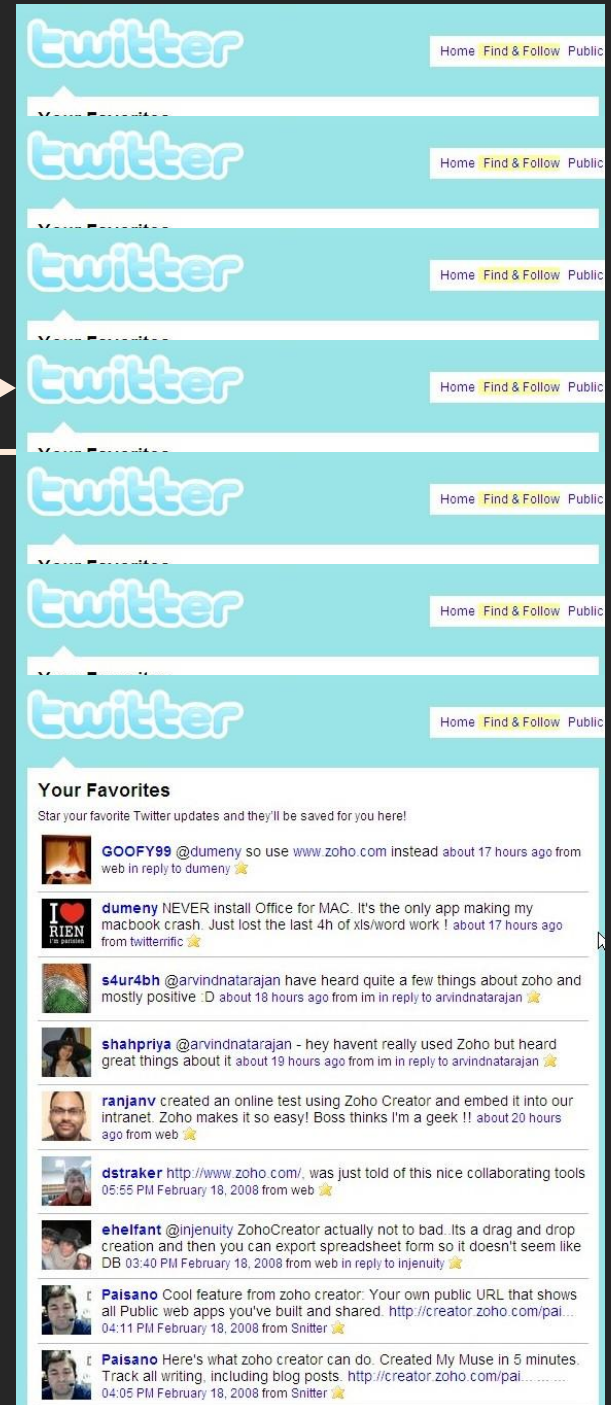
Open Questions

Incorporate
model of tie
relevance in
prediction task?

Learn optimal
threshold for
known feature,
test on unknown
feature?

Question II

- With Scott Counts and Mary Czerwinski, during internship at Microsoft Research, summer 2010



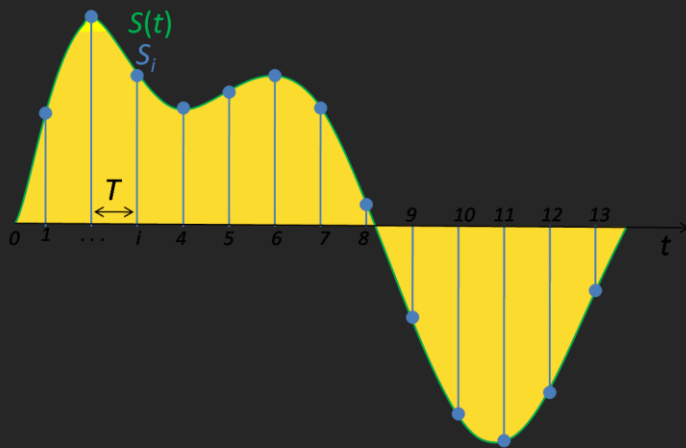
“Information overload”
problem – *Get me the
right content!*



How do we identify the most “relevant” or “best” items on a topic, from millions and even billions of units of social media content?

Let's contrast
this with a
familiar
example


Discrete, regular and fixed sampling lattice



- Shannon-Nyquist sampling theorem: “If a function $x(t)$ contains no frequencies higher than B hertz, it is completely determined by giving its ordinates at a series of points spaced $1/(2B)$ seconds apart.”

Time to sample
each pixel is
constant

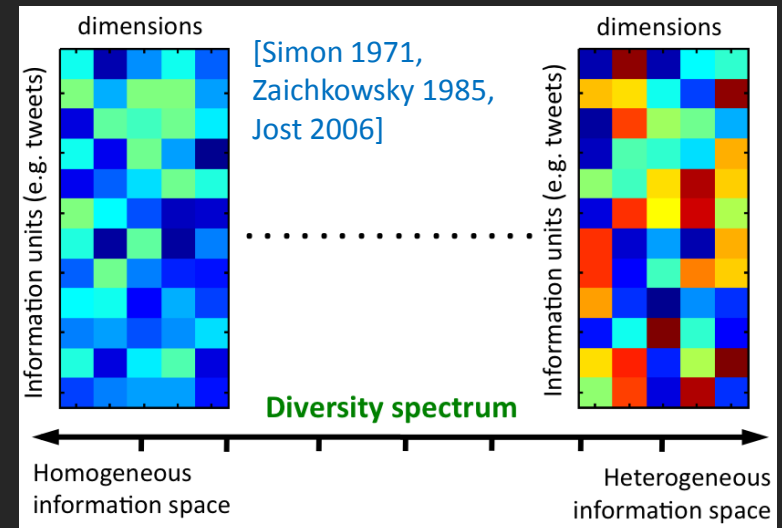
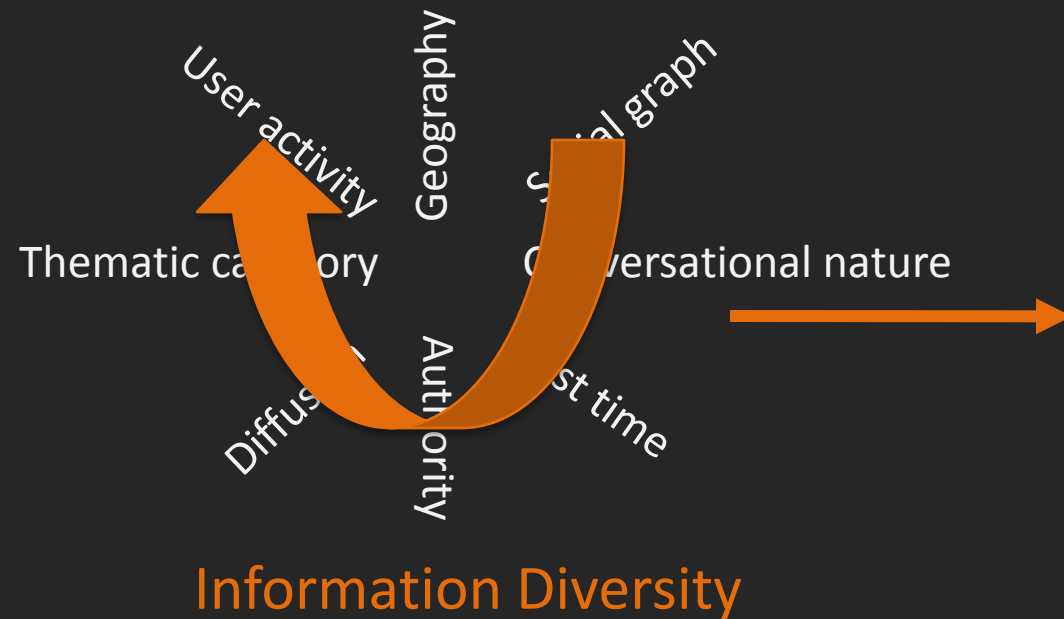
Note that the
web activity has
no notion of
bandwidth!



Interfaces / tools	#Responses
Twitter website	50
Twitter clients, such as Tweetdeck, Twitterific etc.	25
Search engines, such as Bing Social	19
Third party apps, such as Twitter plugin for Google	9

Uni-dimensional
information presentation;
but social media
information is *diverse*.

Characteristics of social media – high dimensionality



Also, social media
sampling needs to
benefit from
mechanisms of
human cognition

“Goodness of a sample” – using measures of human information processing

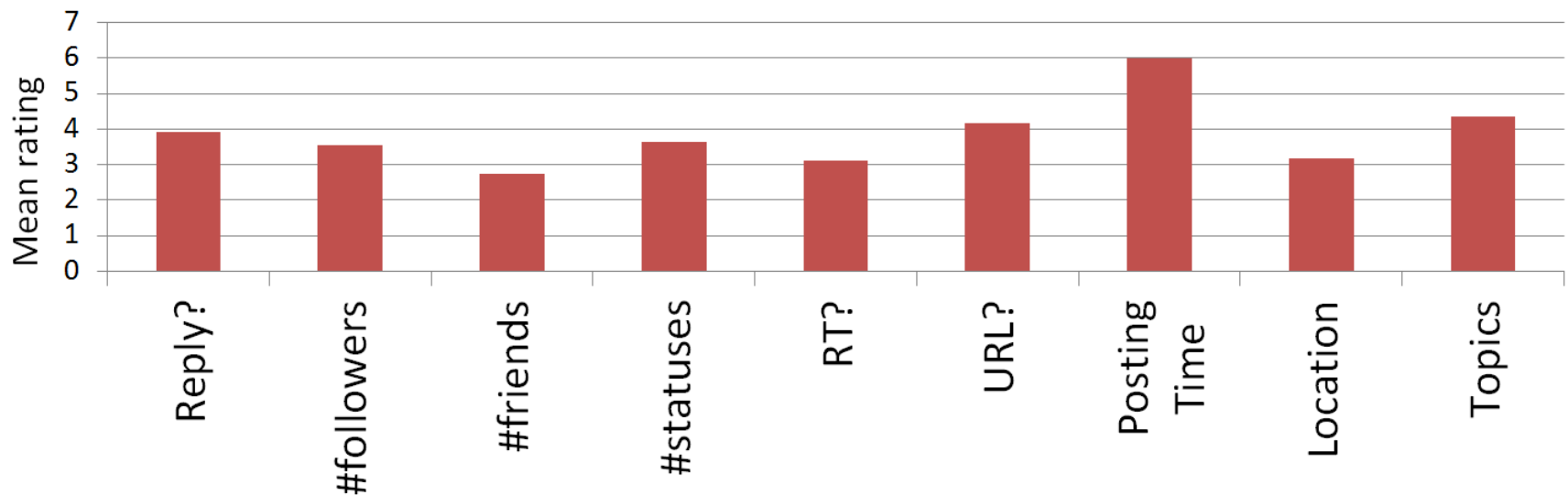


Engagement
Memory encoding
Interestingness
Informativeness

1. What are the significance of various dimensions of social media content?
2. How do we sample such content that matches a certain degree of information diversity?

Dimensional Importance

- Survey based feedback on the importance of different dimensions
 - referred to as “concentration parameters”.
 - Participants (11 ‘active’ Twitter users) were requested to rate each of the tweet dimensions on a scale of 1 through 7, where 1 implied “not important at all”, and 7 meant “highly important”.
 - The survey also allowed them to identify other dimensions that they might think to be significant.

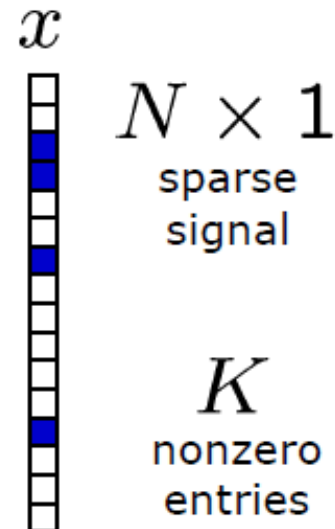


Social media sampling

- Our solution is motivated by the work in the signal processing literature on “compressive sensing” [Candes 2006]:
 - Social media content over time can be considered as signals that often bear the property of being highly “sparse” [Romberg 2008].
 - Compressive sensing can be used to exploit this notion of sparsity in social media content based signals to describe it (i.e. a tweet stream) as a linear combination of a very small number of basis components.

Social media sampling

- Given $x \in \mathbb{R}^{N \times 1}$, we are interested in the “underdetermined” case $M \ll N$, M is the number of basis functions whose coefficients can reconstruct Ψ .
 - Formally, our goal is to find $y \in \mathbb{R}^{M \times 1}$, i.e. the general problem of reconstructing $x \in \mathbb{R}^{N \times 1}$ from linear measurements y about x of the form: $y = \Phi x$, Φ is the transformation matrix.



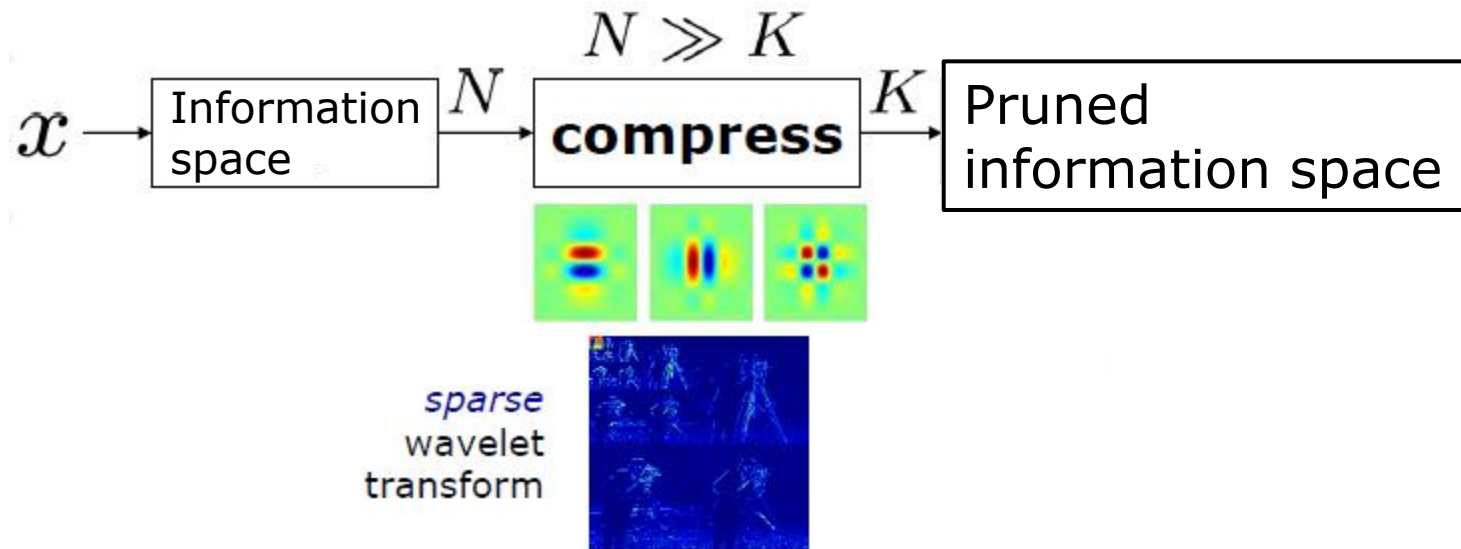
Social media sampling

- Given $x \in \mathbb{R}^{N \times 1}$, we are interested in the “underdetermined” case $M \ll N$, M is the number of basis functions whose coefficients can reconstruct Ψ .
 - Formally, our goal is to find $y \in \mathbb{R}^{M \times 1}$, i.e. the general problem of reconstructing $x \in \mathbb{R}^{N \times 1}$ from linear measurements y about x of the form: $y = \Phi x$, Φ is the transformation matrix.

$$\begin{array}{c} y \\ M \times 1 \\ \text{measurements} \end{array} = \begin{array}{c} \Phi \\ M \times N \end{array} \begin{array}{c} x \\ N \times 1 \\ \text{sparse} \\ \text{signal} \\ K \\ \text{nonzero} \\ \text{entries} \end{array}$$
$$K < M \ll N$$

Social media sampling (contd.)

- We utilize the popular wavelet transform, called “Haar wavelet” for reconstruction of Φ .



Social media sampling (contd.)

- We perform iterative clustering for tweet sample generation – based on entropy distortion minimization technique.
 - The samples are constructed given a sampling ratio ρ and a diversity parameter value ω .
 - The (sub)-optimal sample to be constructed is represented as, $\Psi_s^*(\rho, \omega)$.
- Start with a random tweet as a *sample seed*.
- Iteratively keep on adding tweets from Ψ_s , say t_i , such that the distortion (in terms of L_1 -norm) of entropy of the sample (say, $\Psi_s(i, \omega)$) on addition of the tweet t_i is least with respect to the specified diversity measure ω .

$$\arg \min_{t_i \in \Psi_s, t_i \notin \Psi_s(i-1, \omega)} \|H_o(\Psi_s(i, \omega)) - \omega\|_{L_1}, \text{ where}$$

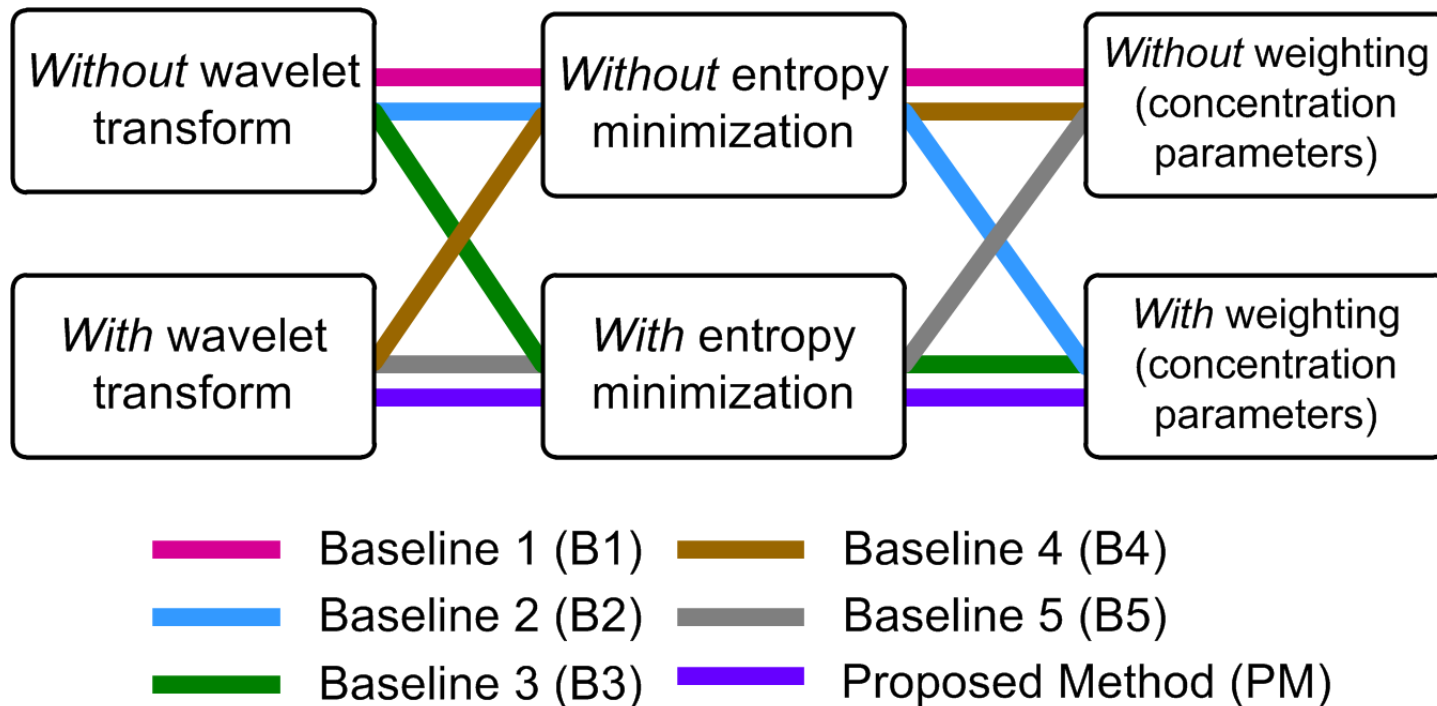
$$H_o(\Psi_s(i, \omega)) = - \sum_{k=1}^K P(\vec{t}_{ik}) \cdot \log P(\vec{t}_{ik}) / H_{\max}, \quad t_i \in \Psi_s \text{ and } H_{\max} = \ln K.$$

How does this method compare to state-of-the-art techniques?

Twitter, **full Firehose**, June 2010, total 1.4 Billion tweets

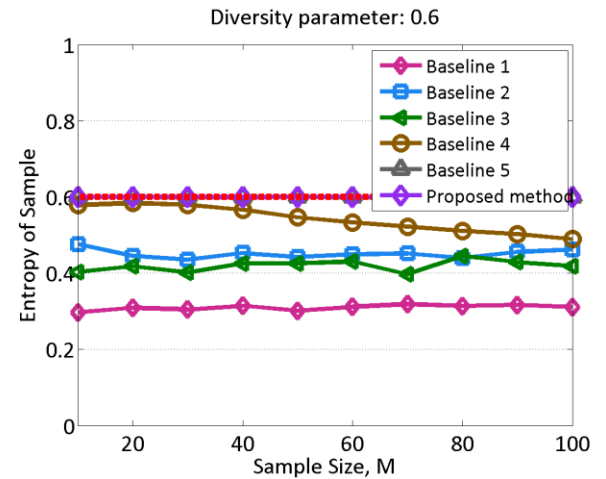
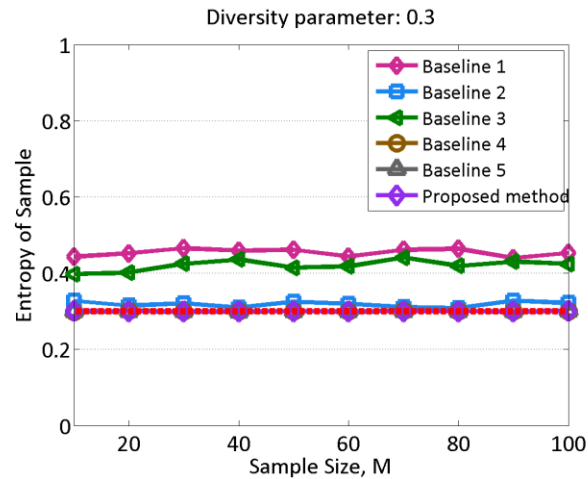
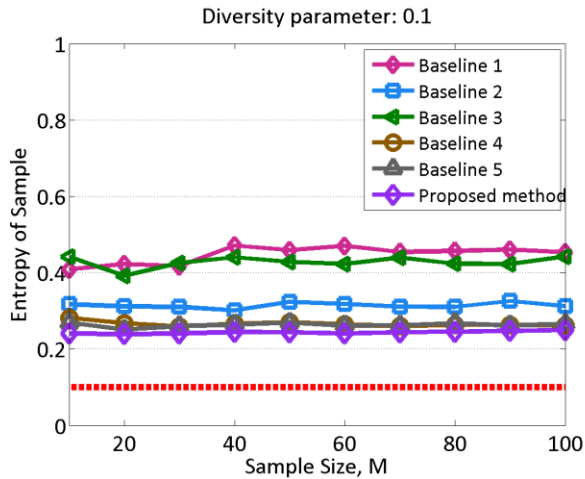
Quantitative evaluation framework

We defined a set of baseline techniques using simplified version of the three different components of our proposed algorithm: use of compression (using wavelet), minimization of entropy and weighting of attributes

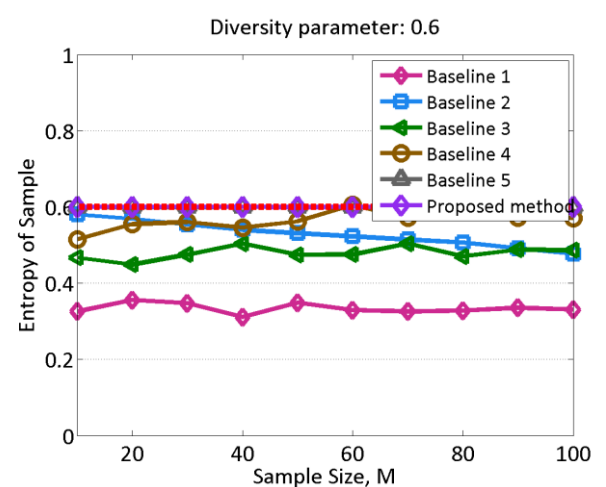
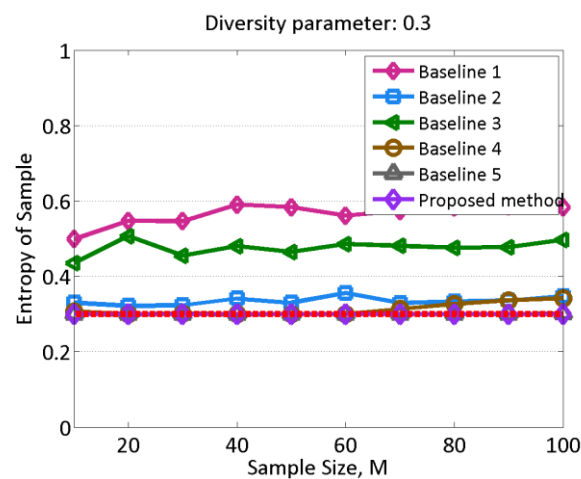
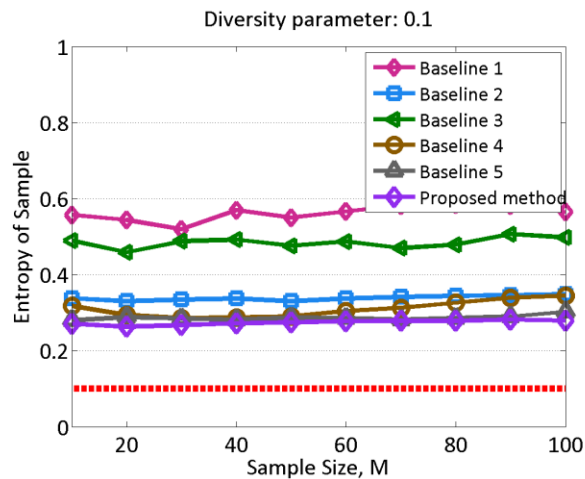


Another two methods: “most recent” tweets and “most tweeted URL” meaning the tweets corresponding to URLs that were highly shared in the network

Quantitative evaluation



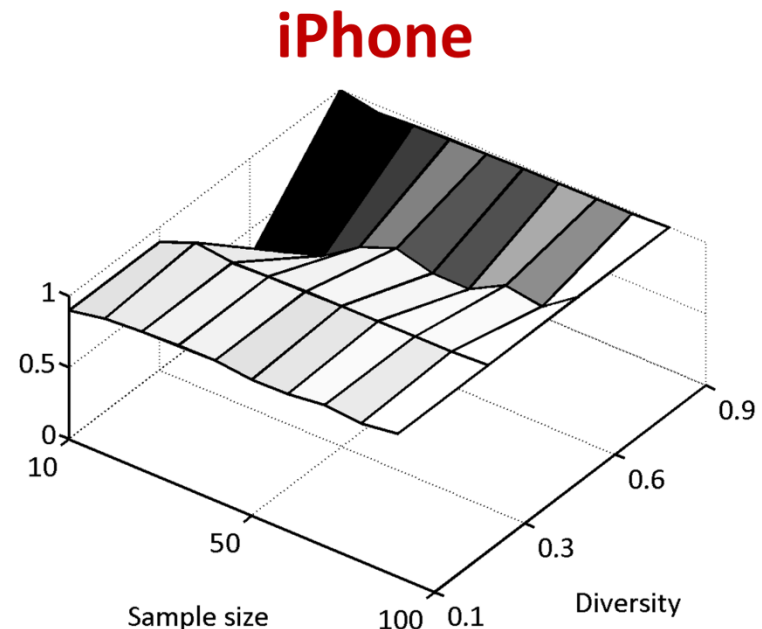
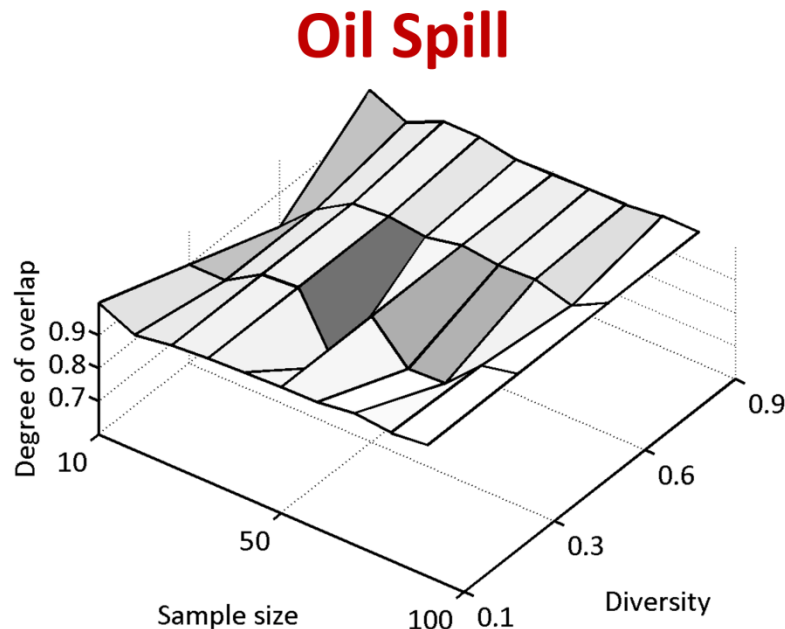
Oil Spill



iPhone

Robustness of sampling method

- Robustness of proposed sampling method across multiple iterations.
 - We show the degree of overlap of tweets corresponding to samples that are generated across iterations. The overlap values are shown for various sample sizes as well as three diversity parameter levels.



How does the sampling
process impact users'
cognitive abilities of
information
consumption?

Cognitive metrics

- *Explicit Measures*. Explicit measures consisted of three 7-point Likert scale ratings made after reading each tweet set,
 - “interestingness”
 - “informativeness”
- *Implicit Measures*.
 - Subjective Duration Assessment [Czerwinski 2001] – ideally if the information presented in a tweet sample is very engaging, the participant would underestimate the time taken to go through it.
 - Recognition Memory for tweets already shown – related to encoding in the long-term memory [Sperling 1973, Smith 1979].

Part I

Please read the following sample of 10 tweets. When you are done reading, click the "Finished Reading!" button below to take a short evaluation of the tweet sample.

Topic: Oil Spill [Tweet Sample, 3 of 12]

From user, @expertox:	Tweet: Will The Oil Spill Affect You? http://blog.expertox.com http://bit.ly/9xt5Od	Posted at: 2010-06-07 06:59:50
From user, @BethanyFilish:	Tweet: RT @rbndvd Blood used to be thicker than water. That was before the BP oil spill though.	Posted at: 2010-06-07 07:00:50
From user, @mattvoss:	Tweet: RT @AP: AP Essay: Gulf oil spill is a reminder of why Americans have lost faith in nearly every national institution. http://bit.ly/cBcK ...	Posted at: 2010-06-07 07:01:24
From user, @chrisjones:	Tweet: http://bit.ly/bpaQD2 Gulf oil spill: Containment cap working well so far, says BP	Posted at: 2010-06-06 15:07:37
From user, @bfergumphs:	Tweet: RT @ScottBourne: If you find this meaningful I'd appreciate a RT - Don't Think Photography's Important? Impact of BP Oil Spill - http:// ...	Posted at: 2010-06-07 06:36:37
From user, @FinanceBreaking:	Tweet: BP Tries To Spin Oil Spill - Watch BP's New Ad (Video) - IndyPosted http://bit.ly/c4kkYQ	Posted at: 2010-06-06 15:40:05
From user, @pinkpanther74:	Tweet: RT @TEDchris: A Gulf oil spill picture I will never forget. http://twitpic.com/1toz8a	Posted at: 2010-06-07 06:43:13
From user, @TheGlobeNews:	Tweet: [The Huffington Post] New Orleans Saints To Visit Oil Spill Areas: Mentions Vince Lombardi Trophy and Bobby Jindal http://fga.me/99fc69	Posted at: 2010-06-06 18:51:51
From user, @Bee_Fly:	Tweet: Oil Spill: http://www.aquarianadvertising.com/info/wordpress/?p=3530	Posted at: 2010-06-07 05:53:45
From user, @1234567890:	Tweet: Oh yeah... Totally forgot about the stupid oil spill. Now I can't swim to the Bahamas lol	Posted at: 2010-06-06 20:20:56

User Study...

Part I

Please read the following sample of 10 tweets. When you are done reading, click the "Finished Reading!" button below to take a short evaluation of the tweet sample.

Topic: Oil Spill [Tweet Sample, 3 of 12]

From user, @expertox:	Tweet: Will The Oil Spill Affect You? http://blog.expertox.com http://bit.ly/9xt5Od	Posted at: 2010-06-07 06:59:50
From user, @BethanyFilish:	Tweet: RT @rbndvd Blood used to be thicker than water. That was before the BP oil spill though.	Posted at: 2010-06-07 07:00:50
From user, @mattvoss:	Tweet: RT @AP: AP Essay: Gulf oil spill is a reminder of why Americans have lost faith in nearly every national institution. http://bit.ly/cBcK ...	Posted at: 2010-06-07 07:01:24
From user, @brianjones:	Tweet: http://bit.ly/bpaQD2 Gulf oil spill: Containment cap working well so far, says BP	Posted at: 2010-06-06 15:07:37
From user, @bfergumphs:	Tweet: RT @ScottBourne: If you find this meaningful I'd appreciate a RT - Don't Think Photography's Important? Impact of BP Oil Spill - http:// ...	Posted at: 2010-06-07 06:36:37
From user, @FinanceBreaking:	Tweet: BP Tries To Spin Oil Spill - Watch BP's New Ad (Video) - IndyPosted http://bit.ly/c4kkYQ	Posted at: 2010-06-06 15:40:05
From user, @pinkpawls:	Tweet: RT @TEDchris: A Gulf oil spill picture I will never forget. http://twitpic.com/1toz8a	Posted at: 2010-06-07 06:43:13
From user, @TheGlobeNewswire:	Tweet: [The Huffington Post] New Orleans Saints To Visit Oil Spill Areas: Mentions Vince Lombardi Trophy and Bobby Jindal http://fga.me/99fc69	Posted at: 2010-06-06 18:51:51
From user, @Beechey:	Tweet: Oil Spill: http://www.aquarianadvertising.com/info/wordpress/?p=3530	Posted at: 2010-06-07 05:53:45
From user, @1234567890:	Tweet: Oh yeah... Totally forgot about the stupid oil spill. Now I can't swim to the Bahamas lol	Posted at: 2010-06-06 20:20:56

Now please respond to the following questions below:

- a. Estimate the length of time, in minutes and seconds (e.g. in the format "X min, Y sec"), you think you needed to go through the tweets.

min, sec
- b. **INTERESTINGNESS:** How interesting did you find the tweets in the sample shown? In the scale below, 1 means not at all interesting, 7 means highly interesting.

1 2 3 4 5 6 7
- c. **DIVERSITY:** How **diverse** did you find the tweets in the sample shown? A diverse set of tweets would contain different sub-topics, would appear to come from different parts of the world, would contain a mix of tweets and re-tweets, etc. In the scale below, 1 means the tweets are not at all diverse, 7 means they are highly diverse.

1 2 3 4 5 6 7
- d. **INFORMATIVENESS:** How informative did you find the tweets in the sample shown? Note, although you'll notice that there are some repeating tweets across samples, rate the informativeness of the sample as a whole. In the scale below, 1 means the sample is not at all informative, and 7 means the sample is highly informative.

1 2 3 4 5 6 7

User Study...

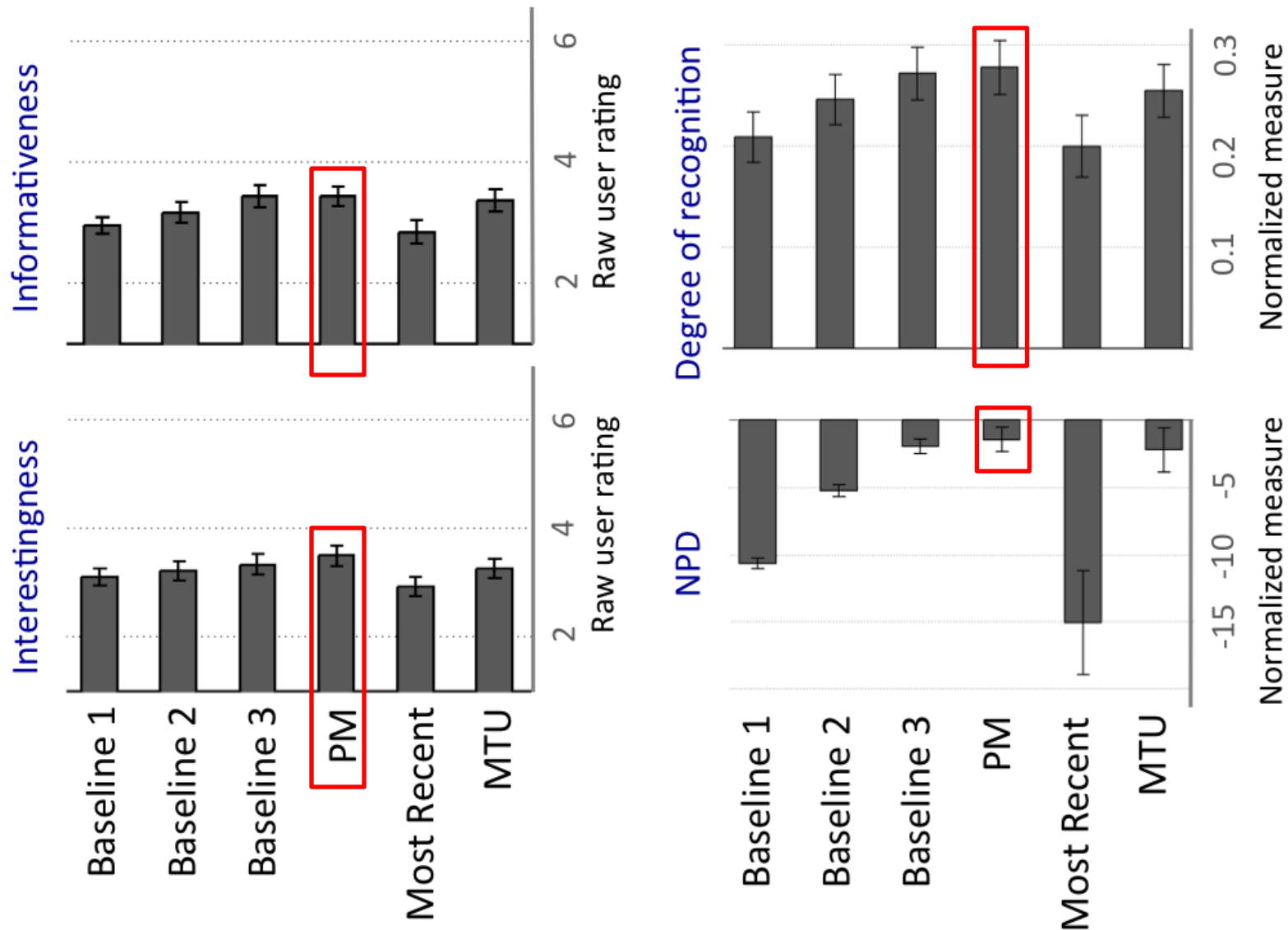
Part III

In this final part of the study you are required to go through the following 72 tweets as presented below. Some of these you would have seen before, while others you wouldn't have seen. Recognize if each of them was shown to you in any of the former pages. Each tweet has a "Yes" / "No" option: so please use your memory to recognize if you saw the tweet or not ("Yes" if you saw it, and "No" if you didn't). Good luck!

From user, @PROBBS	Tweet: RT @malloryallyce: Yo, everyone buy Dawn dish soap \$1 of each bottle goes to helping the poor animals affected by the oil spill. :(Posted at: 2010-06-06 16:59:30	<input type="radio"/> Yes <input type="radio"/> No
From user, @MMAconsciousness	Tweet: RT @ElevateU: RT @PoliticalTicker: House subcommittee holds hearing on oil spill http://bit.ly/cIMJ4a	Posted at: 2010-06-07 06:28:32	<input type="radio"/> Yes <input type="radio"/> No
From user, @misesmises:	Tweet: I think Obama is really killing his chance of re-election with the happening and handling of the BP oil spill. Is this Obama's 9/11?	Posted at: 2010-06-07 16:18:11	<input type="radio"/> Yes <input type="radio"/> No
From user, @BPS747K:	Tweet: RT @nytimesscience: Pelicans, Back from Brink of Extinction, Face Threat From Oil Spill http://nyti.ms/cFGUoN	Posted at: 2010-06-07 12:55:47	<input type="radio"/> Yes <input type="radio"/> No
From user, @fawcettgarnett:	Tweet: [The Huffington Post] New Orleans Saints To Visit Oil Spill Areas: Mentions Vince Lombardi Trophy and Bobby Jindal http://fga.me/99fc69	Posted at: 2010-06-06 18:51:51	<input type="radio"/> Yes <input type="radio"/> No
From user, @MCD2000	Tweet: RT @JasonLeopold: RT @EnvironUpdates: NPR: Scientists: Dispersants Compounded Oil Spill http://bit.ly/dC0V6t Full http://n.pr/b51MvU	Posted at: 2010-06-07 02:44:50	<input type="radio"/> Yes <input type="radio"/> No

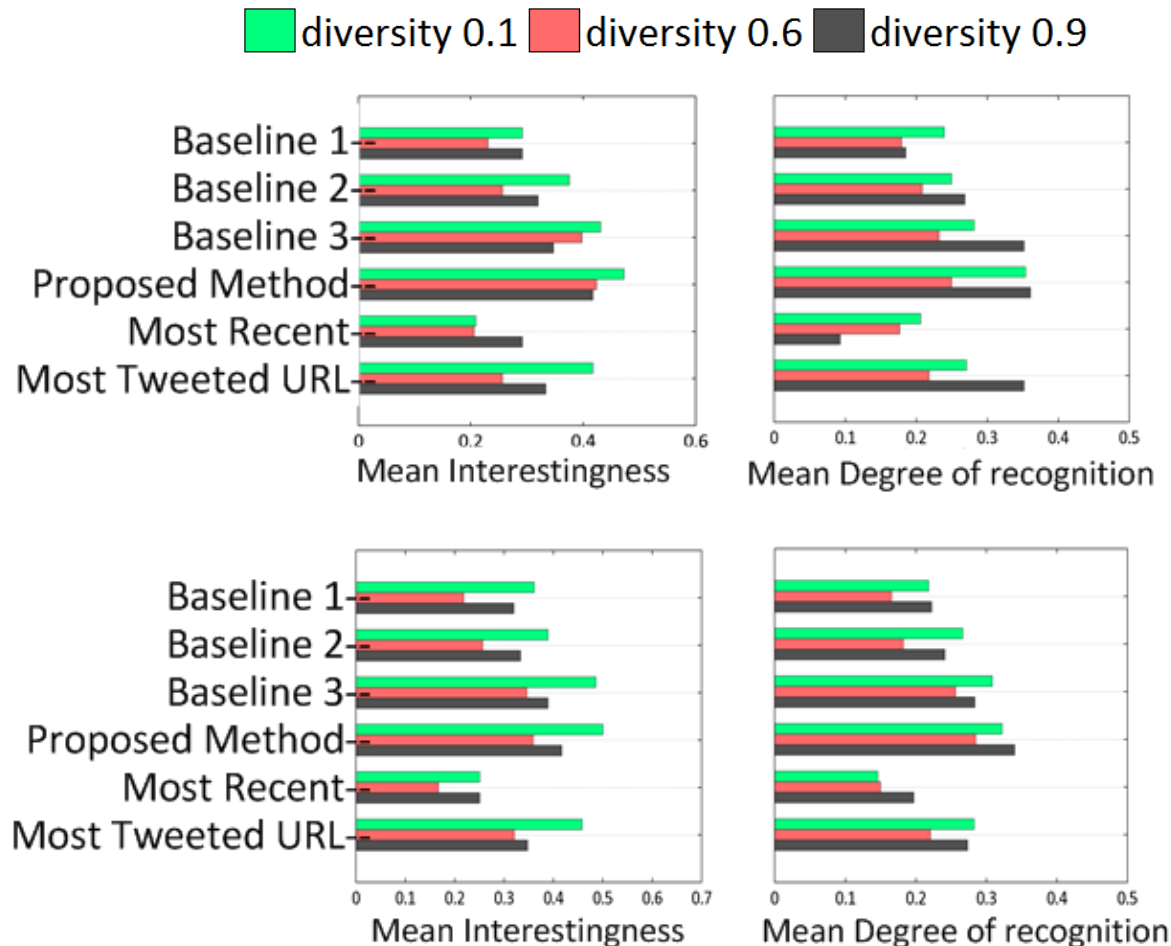
User Study...

Evaluation in terms of Cognitive Metrics



What is the role of
diversity in the sampling
process?

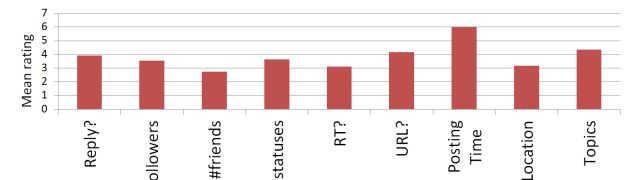
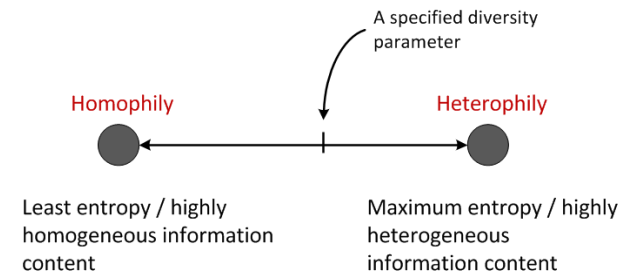
Diversity perception



Participant ratings on different cognitive aspects of information consumption seems to be higher for highly homogenous and highly heterogeneous information samples

Conclusions

- Sampling methodologies of large social information spaces that incorporate cognitive metrics of content consumption can enable the design of better content exploration interfaces.
 - Information diversity is key
 - User appear to cognitively encode information better, when presented with samples of high or low diversity
 - Our proposed sampling algorithms that incorporate cognitive metrics of content consumption perform better than straw-man versions of state-of-the-art techniques



Part I

Please read the following sample of 10 tweets. When you are done reading, click the "Finished Reading!" button below to take a short evaluation of the tweet sample.

Topic: Oil Spill [Tweet Sample, 3 of 12]

From user: @gauravdatta	Tweet: Will The Oil Spill Affect You? http://blog.sagepub.com/http://bit.ly/9d45Qd	Posted at: 2010-06-07 06:59:50
From user: @gauravdatta	Tweet: RT @brenda: Blood used to be thicker than water. That was before the BP oil spill though.	Posted at: 2010-06-07 07:00:50
From user: @gauravdatta	Tweet: RT @AP: AP Essay: Gulf oil spill is a reminder of why Americans have lost faith in nearly every national institution. http://bit.ly/c4ck...	Posted at: 2010-06-07 07:01:24
From user: @gauravdatta	Tweet: http://bit.ly/9d45Qd Gulf oil spill: Containment cap working well so far, says BP	Posted at: 2010-06-06 15:07:37
From user: @gauravdatta	Tweet: RT @ScottBourne: If you find this meaningful I'd appreciate a RT - Don't Think Photography's Important? Impact of BP Oil Spill - http://...	Posted at: 2010-06-07 06:36:37
From user: @gauravdatta	Tweet: BP Tries To Spin Oil Spill - Watch BP's New Ad [Video] - IndyPosted http://bit.ly/c4ckYQ	Posted at: 2010-06-06 15:40:05
From user: @gauravdatta	Tweet: RT @TEDbets: A Gulf oil spill picture I will never forget. http://bit.ly/c4ckYQ	Posted at: 2010-06-07 06:43:13
From user: @gauravdatta	Tweet: [The Huffington Post] New Orleans Saints To Visit Oil Spill Areas: Mentions Vince Lombardi Trophy and Bobby Jindal http://tfg.me/99f6d9	Posted at: 2010-06-06 18:51:51
From user: @gauravdatta	Tweet: Oil Spill: http://www.aquarianadvertising.com/info/wordpress/?p=3530	Posted at: 2010-06-07 05:53:45
From user: @gauravdatta	Tweet: Oh yeah... Totally forgot about the stupid oil spill. Now I can't swim to the Bahamas lol	Posted at: 2010-06-08 20:20:56

Open Questions

Are there empirical
bounds on what
degrees of diversity in a
sample best suit content
consumption?

Does the information
space seem to exhibit
entropy signatures?

If so, can these entropy signatures guide the sampling methodology more adequately and efficiently?

The End

NEXT EXIT 

Social networks
and media are
causing significant
changes in our lives

Inferences
about social
phenomena is
affected by
data quality

Streamlining
the user
experience is
affected by
data relevance

And it matters
completely...



Future Research Directions

Future Directions (Short-term)

- **Social media marketing.**
 - Where (information, people) should one tap a social network to get the optimal/desired effect (economic, technological, cultural) they want?
 - How can our knowledge of the structure of user generated content impact computational advertising on the Web?
- **Study of macroscopic network dynamics from microscopic interactions.**
 - How does (popular) culture evolve on online social systems?
 - How do we characterize emergent order in network amidst noisy communication?

Future Directions (Long-term)

- **Attribute-rich Peta(byte)-scale information spaces.**
 - Social computing and HCI on the cloud
- **Meta-data standards.**
 - What are standards and evaluation metrics that can be developed to generalize communication dynamics over million order data?
- **A comprehensive theory of online communication.**
 - How does such online communication unfold over next-generation systems on the cloud?

Acknowledgements

- HariSundaram, CS +AME, Arizona State University (advisor).
- Doree Duncan Seligmann, Avaya Labs Research.
- Duncan Watts, Yahoo! Research.
- Winter Mason, Yahoo! Research.
- Jake Hofman, Yahoo! Research.
- Scott Counts, Microsoft Research.
- Mary Czerwinski, Microsoft Research.



Questions?

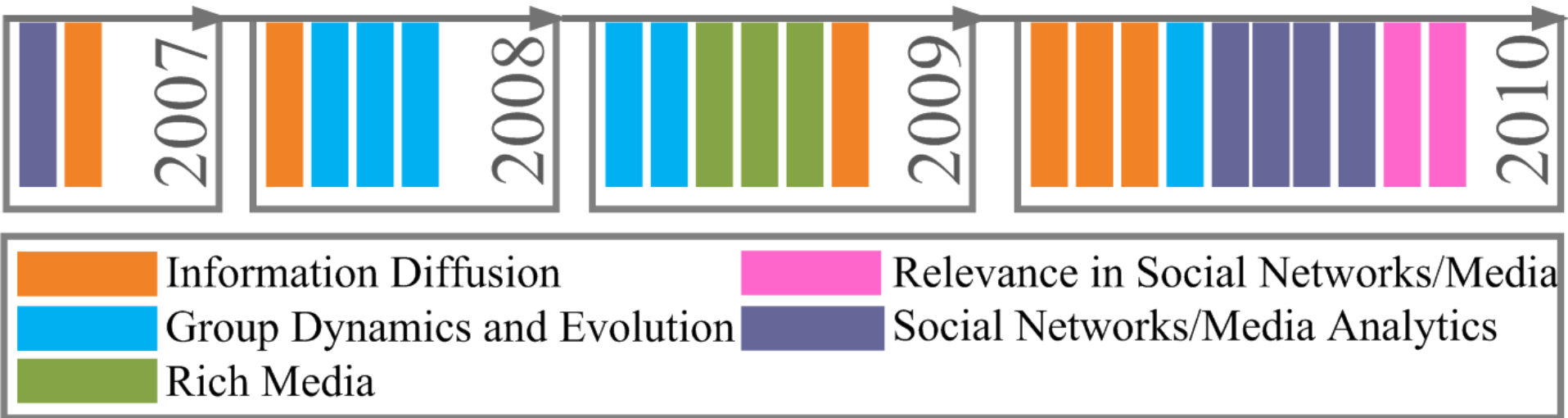
For details: munmun@asu.edu

Web: <http://www.public.asu.edu/~mdechoud/>

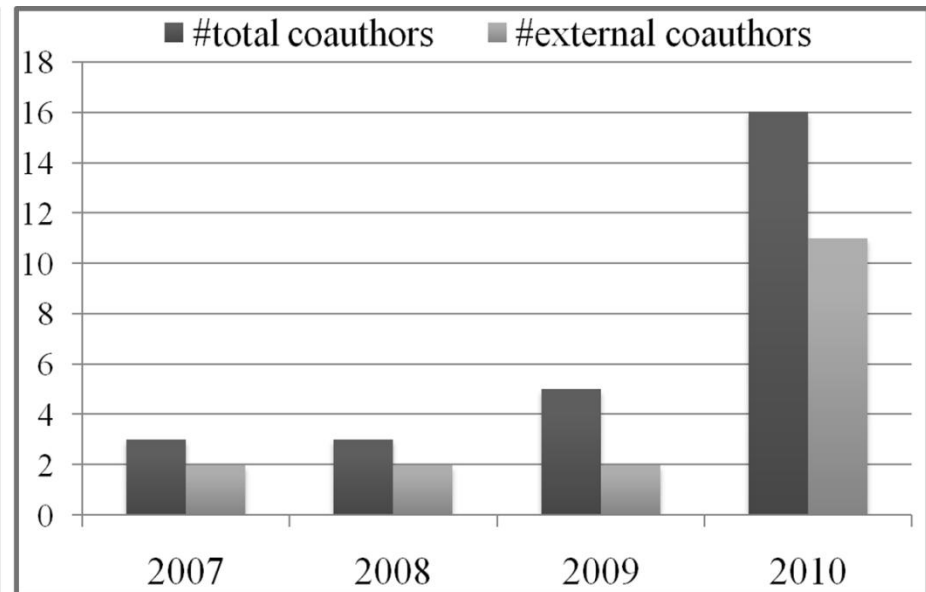
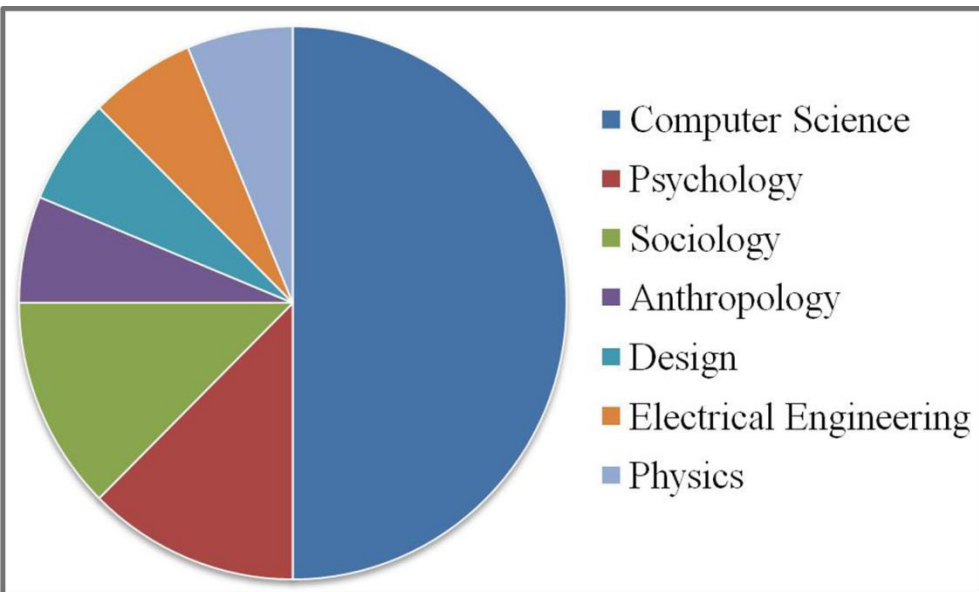
Twitter: @munmun10

Appendix

Topical areas / publications



Interdisciplinary collaboration



Caveats

- Initial assumptions made on social graph construction:
 - Elimination of out-of-network nodes, focusing on a consistent user set over time
 - Geometric mean: alternative definitions of an edge?
 - Considered symmetric edges: communication is often asymmetrical
- Only tested with email datasets
- Type of prediction tasks constrained by available data
- Thresholds on edge weights are not the only way to define edges

Qualitative evaluation

@Paramedic_Fla	Some oil spill events from Monday, June 7, 2010 http://bit.ly/cRwfXn
@miamiauto	Some oil spill events from Monday, June 7, 2010: A summary of events on Monday, June 7, Day 48 of the Gulf of Mexi... http://bit.ly/9HNG9Z
@franklanguage	RT @DAYLEE F@CK that! Broken pipe is not NATURAL! RT @RayBeckermanFreedomWorks CEO, Calls Oil Spill Natural Disaster http://bit.ly/coUY4I
@Teasdallqrb	Public offers 'helpful' ideas on containing BP oil spill - NEWS.com.au

[Twitter search-alike] Most Recent tweets

@_paigenesss	RT @TEDchris: A Gulf oil spill picture I will never forget. http://twitpic.com/1toz8a
@LeiaOfAlderaan	Citizen Speaks The Truth ON BP Gulf Oil Spill--the Govt, BP Are Doing Nothing, There Are No Leaders Here http://bit.ly/BP-Gulf-Oil-Spill
@Faustinagwlxo	WOOW! NO WAY! so brutal! http://ilil.me/h MTV Movie Summer Jam WWDC Oil Spill Xtina Another Cinderella Story
@minxdeluxe	RT @OliBarrett: Visualizing the BP Oil Spill http://www.ifitwasmymhome.com/

[Bing-alike] Most tweeted URL-containing tweets

@JosephAGallant	Erin Brockovich to meet with fishermen who say oil spill dispersant used by BP made them sick. http://huff.to/aGVWII #tcot #BP #oilspill
@dixie_patriot	Oil spill cap catching about 10,000 barrels a day LONDON ? BP's oil spill cap, designed to stop a huge leak from .. http://oohja.com/xeWhD
@MoCuad	My heart breaks all over again, every time I'm reminded of the oil spill.
@NFGNL	Looking for Liability in BP's Gulf Oil Spill: White Collar Watch examines the potential criminal and civil liab.. http://nyti.ms/9IUaT

@jameelee	How You Can Volunteer to Clean Up the Gulf of Mexico Oil Spill http://ow.ly/1V3cu
@conchkid	Gulf;Oil Spill Many Federal Judges Have Links To Oil Industry http://bit.ly/9v45UT
@NewsOnGreen	BP Oil Spill: Containment Cap To Be Replaced Next Month http://dlvr.it/1WDZ8
@TrinitySaveNeo	Citizen Speaks The Truth ON BP Gulf Oil Spill--the Govt, BP Are Doing Nothing, There Are No Leaders Here http://bit.ly/BP-Gulf-Oil-Spill

Interaction between Variables

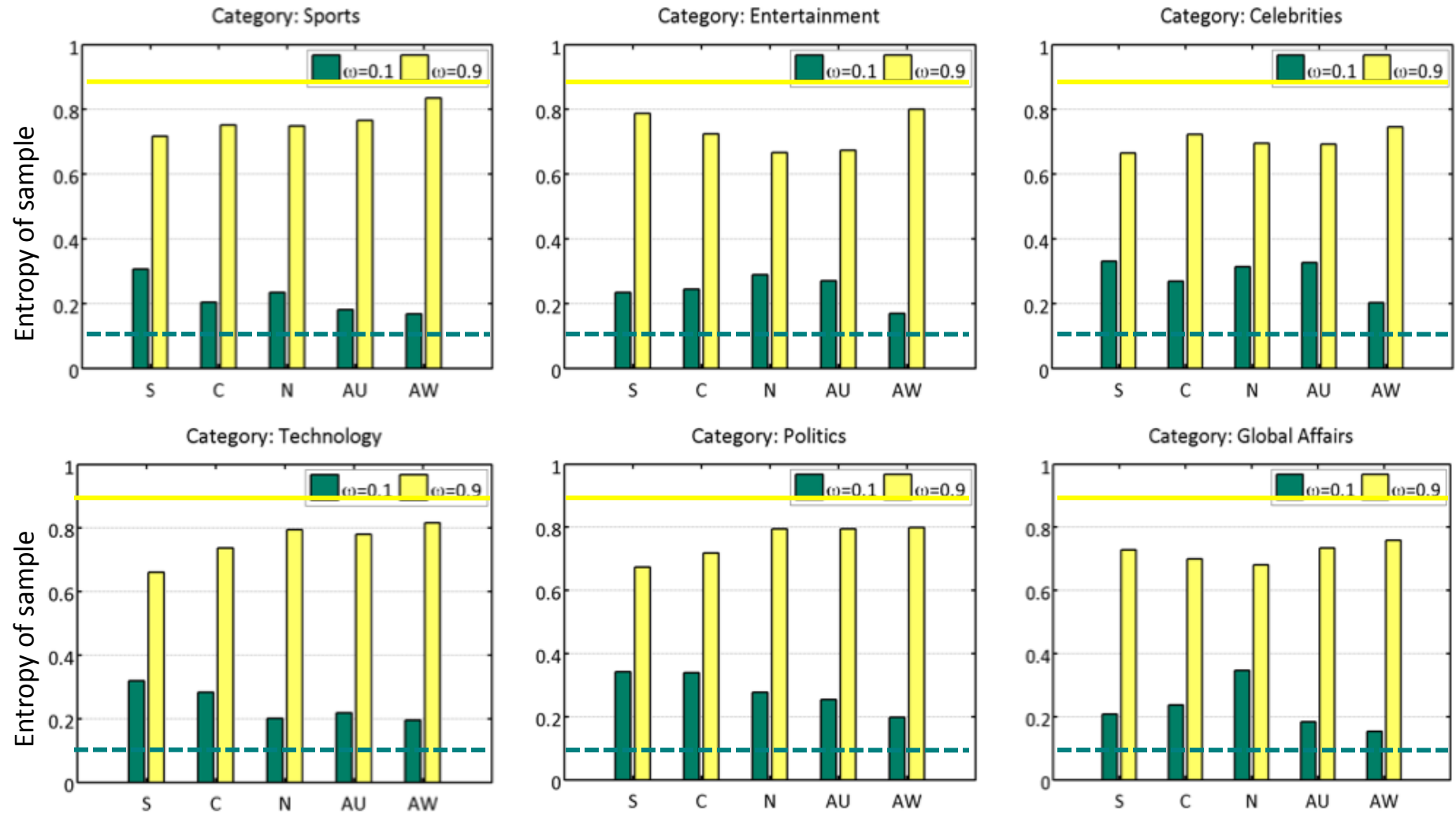
Interaction type	SS	df	MS	F-statistic	p-value
Interestingness					
Diversity × entropy	0.03	5	0.01	0.18	0.9637
Diversity × weighting	0.02	5	0.02	0.08	0.9984
Entropy × weighting	0.02	3	0.01	0.09	0.9517
3-ways	0.29	11	0.03	0.47	0.9190
Informativeness					
Diversity × entropy	0.06	5	0.01	0.40	0.8485
Diversity × weighting	0.08	5	0.02	0.51	0.7647
Entropy × weighting	0.01	3	0.01	0.03	0.9983
3-ways	0.21	11	0.02	0.39	0.9533
Diversity perception					
Diversity × entropy	0.86	5	0.17	5.79	<0.0001
Diversity × weighting	0.85	5	0.17	6.30	<0.0001
Entropy × weighting	0.01	3	0.01	0.20	0.8948
3-ways	1.77	11	0.16	4.16	<0.0001
Normalized perceived duration (NPD)					
Diversity × entropy	493.23	5	98.64	0.80	0.5594
Diversity × weighting	332.31	5	66.46	0.63	0.6738
Entropy × weighting	289.65	3	96.55	2.31	0.0936
3-ways	1780.7	11	161.9	1.03	0.5000
Degree of recognition					
Diversity × entropy	0.15	5	0.03	1.66	0.1625
Diversity × weighting	0.15	5	0.03	1.63	0.1763
Entropy × weighting	0.01	3	0.00	0.11	0.9427
3-ways	0.43	11	0.04	1.14	0.3448

Statistical Significance

	INTERESTINGNESS			INFORMATIVENESS			NPD			DEGREE OF RECOGNITION		
	<i>p</i>	<i>t</i>	<i>d</i>	<i>p</i>	<i>t</i>	<i>d</i>	<i>p</i>	<i>t</i>	<i>d</i>	<i>p</i>	<i>t</i>	<i>d</i>
<i>B1 × PM</i>	0.0028	-2.86	7.83	0.0097	-2.39	5.13	0.0074	-2.51	0.45	0.0974	-1.31	2.78
<i>B2 × PM</i>	0.0278	-1.95	6.95	0.1175	-1.19	1.44	0.0104	-2.37	0.45	0.1055	-1.26	0.79
<i>B3 × PM</i>	0.2401	-0.71	3.94	0.3518	-0.38	8.19	0.1386	-1.09	0.49	0.4117	-0.22	0.22
<i>MR × PM</i>	0.0003	-3.59	14.1	<0.0001	-4.28	452.8	0.0031	-2.84	0.51	0.0052	-2.64	1.58
<i>MTU × PM</i>	0.0607	-1.57	6.22	0.1715	-0.96	1.97	0.0041	-2.72	0.63	0.2142	-0.79	0.57

	diversity=0.1			diversity=0.6			diversity=0.9		
	<i>p</i> -value	<i>t</i>	<i>d</i>	<i>p</i> -value	<i>t</i>	<i>d</i>	<i>p</i> -value	<i>t</i>	<i>d</i>
B1 × PM	0.0476	-1.76	4.69	0.0808	-1.44	2.95	0.0431	-1.81	7.65
B2 × PM	0.1246	-1.19	1.78	0.2897	-0.56	1.23	0.1149	-1.24	2.24
B3 × PM	0.2441	-0.71	1.07	0.3961	-0.27	0.37	0.2384	-0.73	1.07

Impact of Dimensions



S=social, C=content, N=nodal, AU=all features (unweighted), AW=all features (weighted)