Analyzing the Dynamics of Communication in Online Social Networks

Munmun De Choudhury, Hari Sundaram, Ajita John and Doree Duncan Seligmann

Abstract This chapter deals with the analysis of interpersonal communication dynamics in online social networks and social media. Communication is central to the evolution of social systems. Today, the different online social sites feature variegated interactional affordances, ranging from blogging, micro-blogging, sharing media elements (i.e. image, video) as well as a rich set of social actions such as tagging, voting, commenting and so on. Consequently, these communication tools have begun to redefine the ways in which we exchange information or concepts, and how the media channels impact our online interactional behavior. Our central hypothesis is that such communication dynamics between individuals manifest themselves via two key aspects: the information or *concept* that is the content of communication, and the *channel* i.e. the media via which communication takes place. We present computational models and discuss large-scale quantitative observational studies for both these organizing ideas. First, we develop a computational framework to determine the "interestingness" property of conversations cented around rich media. Second, we present user models of diffusion of social actions and study the impact of homophily on the diffusion process. The outcome of this research is twofold. First, extensive empirical studies on datasets from YouTube have indicated that on rich media sites, the conversations that are deemed "interesting" appear to have consequential impact on the properties of the social network they are associated with: in terms of degree of participation of the individuals in future conversations, thematic diffusion as well as emergent cohesiveness in activity among the concerned par-

Munmun De Choudhury
Arizona State University, Tempe, e-mail: munmun@asu.edu

Hari Sundaram
Arizona State University, Tempe e-mail: hari.sundaram@asu.edu

Ajita John
Avaya Labs Research, New Jersey e-mail: ajita@avaya.com

Doree Duncan Seligmann
Avaya Labs Research, New Jersey e-mail: doree@avaya.com

ticipants in the network. Second, observational and computational studies on large social media datasets such as Twitter have indicated that diffusion of social actions in a network can be indicative of future information cascades. Besides, given a topic, these cascades are often a function of attribute homophily existent among the participants. We believe that this chapter can make significant contribution into a better understanding of how we communicate online and how it is redefining our collective sociological behavior.

1 Introduction

During the past decade, the advent of the "social Web" has provided considerable leeway to a rich rubric of platforms that promote communication among users on shared spaces. These interpersonal interactions often take place in the pretext of either a shared media e.g. an image (Flickr), a video (YouTube), a 'blog' / 'microblog' (Twitter); or are built across social ties that reflect human relationships in the physical world (Facebook). The resultant impact of the rapid proliferation of these social websites has been widespread. Individuals today, can express their opinions on personal blogs as well as can share media objects to engage themselves in discussion. Right from shopping a new car, to getting suggestions on investment, searching for the next holiday destination or even planning their next meal out, people have started to rely heavily on opinions expressed online or social resources that can provide them with useful insights into the diversely available set of options. Moreover, personal experiences as well as thoughts and opinions on external events also manifest themselves through "memes", "online chatter" or variegated "voting" mechanisms in several peoples blogs and social profiles.

As a positive outcome of all these interactional affordances provided by the online social media and social network sites, a broad podium of opportunities and ample scope have begun to emerge to the social network analysis community. Instead of focusing on longitudinal studies of relatively small groups such as participant observation [31, 16] and surveys [8], researchers today can study social processes such as information diffusion or community emergence at *very large scales*. This is because electronic social data can be collected at comparatively low cost of acquisition and resource maintenance, can span over diverse populations and be acquired over extended time periods. The result is that study of social processes on a scale of million nodes, that would have been barely possible a few years back, is now looming a lot of interest currently [20, 22].

Our broad goal is to study how such online communication today is reshaping and restructuring our understanding of different social processes. Communication is the process by which participating individuals create and share information with one another in order to reach a mutual understanding [6]. Typically communication involves a form of a *channel*, or a media by means of which information, in the form of *concepts* get transmitted from one individual to another. An illustrative example that describes the key ideas in the online communication process is shown in

Figure 1. Note, mass media channels are more effective in creating knowledge of innovations [5], whereas channels promoting social engagement are more effective in forming and changing attitudes toward a new concept, and thus in influencing the decision to adopt or reject a new concept or information[1].

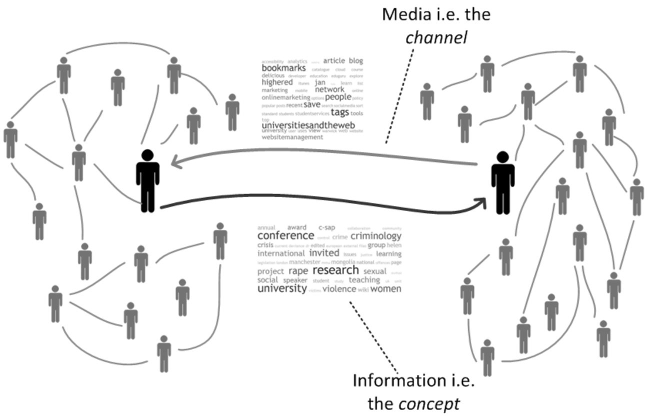

Media i.e. the
channel
accessibility analytics article blog
bookmarks catalogue cloud course
delicious developer education eduguru explore
highered itunes jan learn list
marketing mobile network online
onlinemarketing options people policy
popular posts recent save search socialmedia sort
standard students studentservices tags tools
top
universitiesandtheweb
university user uses view warwick web website
websitemanagement
annual award c-sap collaboration community
conference control crime criminology
crisis current deviance dr edited european external files group helen
international invited issues justice learning
legislation london manchester mmu mongolia national offences page
project rape research sexual
social speaker student study teaching uk unit
university victims violence wiki women
Information i.e.
the *concept*

Fig. 1 Illustration of the two key organizing ideas that embody online interpersonal communication processes: namely, the information or concept that is the content of communication and the channel or the media via which communication takes place.

It, thus, goes without saying that communication is central to the evolution of social systems. To support this empirical finding, over the years, numerous studies on online social communication processes have indicated that studying properties of the associated social system, i.e. the network structure and dynamics can be useful pointers in determining the outcome of many important social and economic relationships [1, 2]. Despite the fundamental importance laid on the understanding of these structures and their temporal behavior in many social and economic settings [8, 10, 9, 20, 21], the development of characterization tools, foundational theoretical models as well as insightful observational studies on large-scale social communication datasets is still in its infancy. This is because communication patterns on online social platforms are significantly distinct from their physical world counterpart—consequently often invalidating the methods, tools and studies designed to cater to longitudinal ethnographic studies on observed physical world interactions. This distinction can be viewed on several aspects relating to the nature of the online communication process itself: such as inexpensive reach to a global audience, volatility of content and easy accessibility of publishing information content online. The outcome of these differences is that today there is an ardent necessity to develop robust computational frameworks to characterize, model and conduct observational studies on online communication processes prevalent, rather pervasively, on the online domain.

[1] Also referred to in popular culture as a "meme".

The contributions of this chapter are also motivated from the potential ability of online communication patterns in addressing multi-faceted sociological, behavioral as well as societal problems. For example, the patterns of social engagement, reflected via the networks play a fundamental role in determining how concepts or information are exchanged. Such information may be as simple as an invitation to a party, or as consequential as information about job opportunities, literacy, consumer products, disease containment and so on. Additionally, understanding the evolution of groups and communities can lend us meaningful insights into the ways in which concepts form and aggregate, opinions develop as well as ties are made and broken, or even how the decisions of individuals contribute to impact on external temporal occurrences. Finally, studies of shared user-generated media content manifested via the communication channel can enable us re-think about the ways in which our communication patterns affect our social memberships or our observed behavior on online platforms.

In the light of the above observations, the following two parts summarize our key research investigations:

- *Rich Media Communication Patterns.* This part investigates rich media communication patterns, i.e. the characteristics of the emergent communication, centered around the *channel* or the shared media artifact. The primary research question we address here is: what are the characteristics of conversations centered around shared rich media artifacts?
- *Information Diffusion.* This part instruments the characterization of the *concept*, or the information or meme, involved in the social communication process. Our central idea encompasses the following question: how do we model user communication behavior that affects the diffusion of information in a social network and what is the impact of user characteristics, such as individual attributes in this diffusion process?

The rest of the chapter is organized as follows. In section 2, we present the major characteristics of online communication dynamics. Next two sections deal with the methods that help us study rich media communication patterns (section 3) and impact of communication properties on diffusion processes (section 4). They also present some experimental studies conducted on large-scale datasets to evaluate our proposed methods of communication analysis. Finally we conclude in section 5 with a summary of the contributions and future research opportunities.

2 Characteristics of Online Communication

We present key characteristics of the online communication process. First we present a background survey of the different aspects of online communication. Next we discuss the different forms of communication affordances that are provided by different online social spaces today and discuss an overview of prior work on the different modalities.

Table 1 Some social media statistics.

SOCIAL MEDIA TYPE	KEY STATISTICS
YouTube	139M users; US$200M [Forbes]
Flickr	3.6B images; 50M users
Facebook	350M active users; 1B pieces of content shared each week
MySpace	110M monthly active users; 14 B comments on the site
Digg	3M unique users; $40M
Engadget	1,887,887 monthly visitors
Huffington Post	8.9M visitors
Live Journal	19,128,882 accounts

2.1 Background

There are several ways in which online social media has revolutionized our means and manner of social communication today: naturally making a huge impact on the characteristics of the social systems that encompass them. We discuss some of the characteristics of this widespread change in the communication process as follows:

1. *Reach.* Social media communication technologies provide scale and enable anyone to reach a global audience.
2. *Accessibility.* Social media communication tools are generally available to anyone at little or no cost, converting every individual participant in the online social interaction into a publisher and broadcaster of information content on their own.
3. *Usability.* Most social media do not, or in some cases reinvent skills, so anyone can operate the means of content production and subsequent communication, eliminating most times the need for specialized skills and training.
4. *Recency.* Social media communication can be capable of virtually instantaneous responses; only the participants determining any delay in response; making the communication process extremely reciprocative, with low lags in responses.
5. *Permanence.* unlike industrial media communication, which once created, cannot be altered (e.g. once a magazine article is printed and distributed changes cannot be made to that same article), social media communication is extremely volatile over time, because it can be altered almost instantaneously by comments, editing, voting and so on.

These key characteristics of online social communication have posed novel challenges on the study of social systems in general. To highlight some of the key statistics of different social sites available on the Web today, we compiled Table 1. The natural question that arises is that: how are online social communication patterns today affecting our social lives and our collective behavior? As is obvious from the statistics, traditional tools to understand social interactions in physical spaces or over industrial media or even prior work involving longitudinal studies of groups of individuals are therefore often only partially capable of characterizing, modeling and observing the modern online communication of today.

In this chapter, we therefore identify two key components that subsume these diverse characteristics of the online social communication process on social media today. These two components are manifested as below:

1. The entity or the concept (e.g. information, or 'meme').
2. The channel or the media (e.g. textual, audio, video or image-based interactive channel).

2.2 Communication Modes in Social Networks

We discuss several different communication modes popularly existent in social networks and social media sites today. These diverse modalities of communication allow users to engage in interaction often spanning a commonly situated interest, shared activities or artifacts, geographical, ethnic or gender-based co-location, or even dialogue on external news events. In this chapter, we have focused on the following forms of communication among users, that are likely to promote social interaction:

1. *Messages.* Social websites such as MySpace feature an ability to users to post short messages on their friends' profiles. A similar feature on Facebook allows users to post content on another user's "Wall". These messages are typically short and viewable publicly to the common set of friends to both the users; providing evidences of interaction via communication.
2. *Blog Comments / Replies.* Commenting and replying capability provided by different blogging websites, such as Engadget, Huffington Post, Slashdot, Mashable or MetaFilter provide substantial evidence of back and forth communication among sets of users, often relating to the topic of the blog post. Note, replies are usually shown as an indented block in response to the particular comment in question.
3. *Conversations around Shared Media Artifact.* Many social websites allow users to share media artifacts with their local network or set of contacts. For example, on Flickr a user can upload a photo that is viewable via a feed to her contacts; while YouTube allows users to upload videos emcompassing different topical categories. Both these kinds of media sharing allow rich communication activity centered around the media elements via comments. These comments often take a conversational structure, involving considerable back and forth dialogue among users.
4. *Social Actions.* A different kind of a communication modality provided by certain social sites such as Digg or del.icio.us involves participation in a variety of social actions by users. For example, Digg allows users to vote (or rate) on shared articles, typically news, via a social action called "digging". Another example is the "like" feature provided by Facebook on user statuses, photos, videos and shared links. Such social action often acts as a proxy for communication activ-

ity, because first, it is publicly observable, and second it allows social interaction among the users.

5. *Micro-blogging.* Finally, we define a communication modality based on micro-blogging activity of users, e.g. as provided by Twitter. The micro-blogging feature, specifically called “tweeting” on Twitter, often takes conversational form, since tweets can be directed to a particular user as well. Moreover, Twitter allows the “RT” or re-tweet feature, allowing users to propagate information from one user to another. Hence micro-blogging activity can be considered as an active interactional medium.

2.3 Prior Work on Communication Modalities

In this section we will survey some prior work on the above presented communication modalities.

Conversations. Social networks evolve centered around communication artifacts. The conversational structure by dint of which several social processes unfold, such as diffusion of innovation and cultural bias, discovery of experts or evolution of groups, is valuable because it lends insights into the nature of the network at multi-grained temporal and topological levels and helps us understand networks as an emergent property of social interaction.

Comments and messaging structure in blogs and shared social spaces have been used to understand dialogue based conversational behavior among individuals [34] as well as in the context of summarization of social activity on the online platform or to understand the descriptive nature of web comments [32]. Some prior work have also deployed conversational nature of comments to understand social network structure as well as in statistical analysis of networks [15]. There has also been considerable work on analyzing discussions or comments in blogs [28] as well as utilizing such communication for prediction of its consequences like user behavior, sales, stock market activity etc.

Prior research has also discovered value in using social interactional data to understand and in certain cases predict external behavioral phenomena [11]. There has been considerable work on analyzing social network characteristics in blogs [20] as well as utilizing such communication for prediction of its consequences like user behavior, sales, stock market activity etc [3, 17]. In [17] Gruhl et al. attempt to determine if blog data exhibit any recognizable pattern prior to spikes in the ranking of the sales of books on Amazon.com. Adar et al. in [3] present a framework for modeling and predicting user behavior on the web. They created a model for several sets of user behavior and used it to automatically compare the reaction of a user population on one medium e.g. search engines, blogs etc to the reactions on another.

Social Actions. The participation of individual users in online social spaces is one of the most noted features in the recent explosive growth of popular online communities ranging from picture and video sharing (Flickr.com and YouTube.com) and collective music recommendation (Last.fm) to news voting (Digg.com) and social bookmarking (del.icio.us). However in contrast to traditional communities, these sites do not feature direct communication or conversational mechanisms to its members. This has given rise to an interesting pattern of social action based interaction among users. The users' involvement and their contribution through non-message-based interactions, e.g. digging or social bookmarking have become a major force behind the success of these social spaces. Studying this new type of user interactional modality is crucial to understanding the dynamics of online social communities and community monetization.

Social actions [12] performed on shared spaces often promote rich communication dynamics among individuals. In prior work, authors have discussed how the voting i.e. digging activity on Digg impacts the discovery of novel information [37]. Researchers [35] have also examined the evolution of activity between users in the Facebook social network to capture the notion of how social links can grow stronger or weaker over time. Their experiments reveal that links in the activity network on Facebook tend to come and go rapidly over time, and the strength of ties exhibits a general decreasing trend of activity as the social network link ages. Social actions revealed via third party applications as featured by Facebook have also lent interesting insights into the social characteristics of online user behavior.

In this chapter, we organize our approach based on these two different modalities of online communication, i.e. conversations and social actions. We utilize the former to study the dynamic characterization of the media channel that embodies online communication. While the latter is used to study the diffusion properties of the concept or the unit of information that is transmitted in a network via the communication process. This is presented in the following two sections.

3 Rich Media Communication Patterns

An interesting emergent property of large-scale user-generated content on social media sites is that these shared media content seem to generate rich dialogue of communication centered round shared media objects, e.g. YouTube, Flickr etc. Hence apart from impact of communication on the dynamics of the individuals' actions, roles and the community in general, there are additional challenges on how to characterize such "conversations", understanding the relationship of the conversations to social engagement i.e. the community under consideration, as well as studying the observed user behavior responsible for publishing and participation of the content.

Today, there is significant user participation on rich media social networking websites such as YouTube and Flickr. Users can create (e.g. upload photo on Flickr), and consume media (e.g. watch a video on YouTube). These websites also allow for significant communication between the users—such as comments by one user on a media uploaded by another. These comments reveal a rich dialogue structure (user *A* comments on the upload, user *B* comments on the upload, *A* comments in response to *B*'s comment, *B* responds to *A*'s comment etc.) between users, where the discussion is often about themes unrelated to the original video. In this section, the sequence of comments on a media object is referred to as a conversation. Note the theme of the conversation is latent and depends on the content of the conversation.

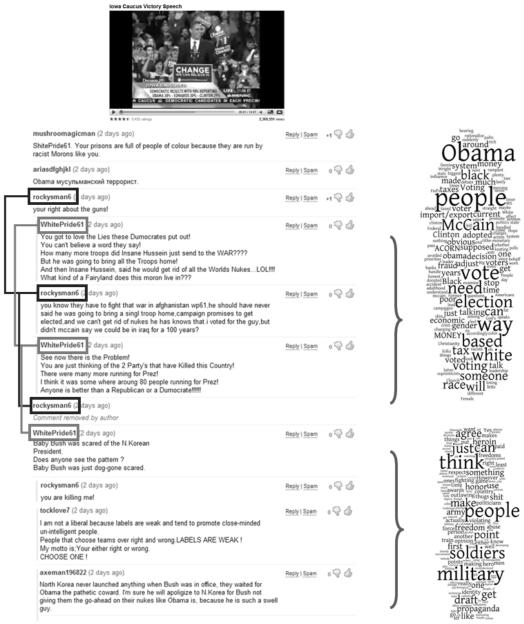
The fundamental idea explored in this section is that analysis of communication activity is crucial to understanding repeated visits to a rich media social networking site. People return to a video post that they have already seen and post further comments (say in YouTube) in response to the communication activity, rather than to watch the video again. Thus it is the content of the communication activity itself that the people want to read (or see, if the response to a video post is another video, as is possible in the case of YouTube). Furthermore, these rich media sites have notification mechanisms that alert users of new comments on a video post / image upload promoting this communication activity.

We denote the communication property that causes people to further participate in a conversation as its "interestingness." While the meaning of the term "interestingness" is subjective, we decided to use it to express an intuitive property of the communication phenomena that we frequently observe on rich media networks. Our goal is to determine a real scalar value corresponding to each conversation in an objective manner that serves as a measure of interestingness. Modeling the user subjectivity is beyond the scope of this section.

What causes a conversation to be interesting to prompt a user to participate? We conjecture that people will participate in conversations when (a) they find the conversation theme interesting (what the previous users are talking about) (b) see comments by people that are well known in the community, or people that they know directly comment (these people are interesting to the user) or (c) observe an engaging dialogue between two or more people (an absorbing back and forth between two people). Intuitively, interesting conversations have an engaging theme, with interesting people. Example of an interesting conversation from YouTube is shown in Figure 2.

A conversation that is deemed interesting must be consequential [13]—i.e. it must impact the social network itself. Intuitively, there should be three consequences (a) the people who find themselves in an interesting conversation, should tend to co-participate in future conversations (i.e. they will seek out other interesting people that they've engaged with) (b) people who participated in the current interesting conversation are likely to seek out other conversations with themes similar to the current conversation and finally (c) the conversation theme, if engaging, should slowly proliferate to other conversations.

There are several reasons why measuring interestingness of a conversation is of value. First, it can be used to rank and filter both blog posts and rich media, particularly when there are multiple sites on which the same media content is posted, guiding users to the most interesting conversation. For example, the same news story may be posted on several blogs, our measures can be used to identify those sites where the postings and commentary is of greatest interest. It can also be used to increase efficiency. Rich media sites, can manage resources based on changing interestingness measures (e.g. and cache those videos that are becoming more in-



Iowa Caucus Victory Speech

mushroomagicman (2 days ago) Reply | Spam +1

ShitePride61. Your prisons are full of people of colour because they are run by racist Morons like you.

ariasdfghjkl (2 days ago) Reply | Spam 0

Obama мусульманский террорист.

rockysman6 (days ago) Reply | Spam +1

your right about the guns!

WhitePride61 2 days ago) Reply | Spam 0

You got to love the Lies these Dumocrates put out!
You can't believe a word they say!
How many more troops did Insane Hussein just send to the WAR????
But he was going to bring all the Troops home!
And then Insane Hussein, said he would get rid of all the Worlds Nukes...LOL!!!!
What kind of a Fairyland does this moron live in???

rockysman6 (days ago) Reply | Spam 0

you know they have to fight that war in afghanistan wp61,he should have never said he was going to bring a singl troop home,campaign promises to get elected,and we can't get rid of nukes he has knows that.i voted for the guy,but didn't mccain say we could be in iraq for a 100 years?

WhitePride61 2 days ago) Reply | Spam 0

See now there is the Problem!
You are just thinking of the 2 Party's that have Killed this Country!
There were many more running for Prez!
I think it was some where aroung 80 people running for Prez!
Anyone is better than a Republican or a Dumocrate!!!!!!

rockysman6 (2 days ago)

Comment removed by author

WhitePride61 (2 days ago) Reply | Spam 0

Baby Bush was scared of the N.Korean
President.
Does anyone see the pattern ?
Baby Bush was just dog-gone scared.

rockysman6 (2 days ago) Reply | Spam 0

you are killing me!

tocklove7 (2 days ago) Reply | Spam 0

I am not a liberal because labels are weak and tend to promote close-minded un-intelligent people.
People that choose teams over right and wrong.LABELS ARE WEAK !
My motto is;Your either right or wrong.
CHOOSE ONE !

axeman196822 (2 days ago) Reply | Spam 0

North Korea never launched anything when Bush was in office, they waited for Obama the pathetic coward. I'm sure he will apoligize to N.Korea for Bush not giving them the go-ahead on their nukes like Obama is, because he is such a swell guy.

Fig. 2 Example of an interesting conversation from YouTube. Note it involves back-and-forth dialogue between participants as well as evolving themes over time.

teresting), and optimize retrieval for the dominant themes of the conversations. Besides, differentiated advertising prices for ads placed alongside videos can be based on their associated conversational interestingness.

3.1 Problem Formulation

3.1.1 Definitions

Conversation. We define a conversation in online social media (e.g., an image, a video or a blog post) as a temporally ordered sequence of comments posted by individuals whom we call "participants". In this section, the content of the conversations are represented as a stemmed and stop-word eliminated bag-of-words.

Conversational Themes. Conversational themes are sets of salient topics associated with conversations at different points in time.

Interestingness of Participants. Interestingness of a participant is a property of her communication activity over different conversations. We propose that an interesting participant can often be characterized by (a) several other participants writing comments after her, (b) participation in a conversation involving other interesting participants, and (c) active participation in "hot" conversational themes.

Interestingness of Conversations. We now define "interestingness" as a dynamic communication property of conversations which is represented as a real non-negative scalar dependent on (a) the evolutionary conversational themes at a particular point of time, and (b) the communication properties of its participants. It is important to note here that "interestingness" of a conversation is necessarily subjective and often depends upon context of the participant. We acknowledge that alternate definitions of interestingness are also possible.

Conversations used in this section are the temporal sequence of comments associated with media elements (videos) in the highly popular media sharing site YouTube. However our model can be generalized to any domain with observable threaded communication. Now we formalize our problem based on the following data model.

3.1.2 Data Model

Our data model comprises the tuple C, P having the following two inter-related entities: a set of conversations, C on shared media elements; and a set of participants P in these conversations. Each conversation is represented with a set of comments, such that each comment that belongs to a conversation is associated with a unique participant, a timestamp and some textual content (bag-of-words).

We now discuss the notations. We assume that there are N participants, M conversations, K conversation themes and Q time slices. Using the relationship between the entities in the tuple C, P from the above data model, we construct the following matrices for every time slice $q, 1 \leq q \leq Q$:

- $\mathbf{P_F}^{(q)} \in \mathbb{R}^{N \times N}$: Participant-follower matrix, where $\mathbf{P_F}^{(q)}(i,j)$ is the probability that at time slice q, participant j comments following participant i on the conversations in which i had commented at any time slice from 1 to $(q-1)$.
- $\mathbf{P_L}^{(q)} \in \mathbb{R}^{N \times N}$: Participant-leader matrix, where $\mathbf{P_L}^{(q)}(i,j)$ is the probability that in time slice q, participant i comments following participant j on the conversations in which j had commented in any time slice from 1 to $(q-1)$. Note, both $\mathbf{P_F}^{(q)}$ and $\mathbf{P_L}^{(q)}$ are asymmetric, since communication between participants is directional.
- $\mathbf{P_C}^{(q)} \in \mathbb{R}^{N \times M}$: Participant-conversation matrix, where $\mathbf{P_C}^{(q)}(i,j)$ is the probability that participant i comments on conversation j in time slice q.
- $\mathbf{C_T}^{(q)} \in \mathbb{R}^{M \times K}$: Conversation-theme matrix, where $\mathbf{C_T}^{(q)}(i,j)$ is the probability that conversation i belongs to theme j in time slice q.
- $\mathbf{T_S}^{(q)} \in \mathbb{R}^{K \times 1}$: Theme-strength vector, where $\mathbf{T_S}^{(q)}(i)$ is the strength of theme i in time slice q. Note, $\mathbf{T_S}^{(q)}$ is simply the normalized column sum of $\mathbf{C_T}^{(q)}$.
- $\mathbf{P_T}^{(q)} \in \mathbb{R}^{N \times K}$: Participant-theme matrix, where $\mathbf{P_T}^{(q)}(i,j)$ is the probability that participant i communicates on theme j in time slice q. Note, $\mathbf{P_T}^{(q)} = \mathbf{P_C}^{(q)} \cdot \mathbf{C_T}^{(q)}$.

- $\mathbf{I_P}^{(q)} \in \mathbb{R}^{N\times 1}$: Interestingness of participants vector, where $\mathbf{I_P}^{(q)}(i)$ is the interestingness of participant i in time slice q.
- $\mathbf{I_C}^{(q)} \in \mathbb{R}^{M\times 1}$: Interestingness of conversations vector, where $\mathbf{I_C}^{(q)}(i)$ is the interestingness of conversation i in time slice q.

For simplicity of notation, we denote the i-th row of the above 2-dimensional matrices as $\mathbf{X}(i,:)$.

3.1.3 Problem Statement

Now we formally present our problem statement: given a dataset C,P and associated meta-data, we intend to determine the interestingness of the conversations in C, defined as $\mathbf{I_C}^{(q)}$ (a non-negative scalar measure for a conversation) for every time slice $q, 1 \leq q \leq Q$. Determining interestingness of conversations involves two key challenges:

1. How to extract the evolutionary conversational themes?
2. How to model the communication properties of the participants through their interestingness?

Further in order to justify interestingness of conversations, we need to address the following challenge: what are the consequences of an interesting conversation?

In the following three sections, we discuss how we address these three challenges through: (a) detecting conversational themes based on a mixture model that incorporates regularization with time indicator, regularization for temporal smoothness and for co-participation; (b) modeling interestingness of participants; and of interestingness of conversations; and using a novel joint optimization framework of interestingness that incorporates temporal smoothness constraints and (c) justifying interestingness by capturing its future consequences.

3.2 Conversational Themes

In this section, we discuss the method of detecting conversational themes. We elaborate on our theme model in the following two sub-sections—first a sophisticated mixture model for theme detection incorporating time indicator based, temporal and co-participation based regularization is presented. Second, we discuss parameter estimation of this theme model.

3.2.1 Chunk-based Mixture Model of Themes

Conversations are dynamically growing collections of comments from different participants. Hence, static keyword or tag based assignment of themes to conversations independent of time is not useful. Our model of detecting themes is therefore based on segmentation of conversations into ‘chunks’ per time slice. A chunk is a representation of a conversation at a particular time slice and it comprises a (stemmed and stop-word eliminated) set of comments (bag-of-words) whose posting timestamps lie within the same time slice. Our goal is to associate each chunk (and hence the conversation at that time slice) with a theme distribution. We develop a sophisticated multinomial mixture model representation of chunks over different themes (a modified pLSA [18]) where the theme distributions are (a) regularized with time indicator, (b) smoothed across consecutive time slices, and (c) take into account the prior knowledge of co-participation of individuals in the associated conversations.

Let us assume that a conversation c_i is segmented into Q non-overlapping chunks (or bag-of-words) corresponding to the Q different time slices. Let us represent the chunk corresponding to the i-th conversation at time slice $q(1 \leq q \leq Q)$ as $\lambda_{i,q}$. We further assume that the words in $\lambda_{i,q}$ are generated from K multinomial theme models $\theta_1, \theta_2, \cdots, \theta_K$ whose distributions are hidden to us. Our goal is to determine the log likelihood that can represent our data, incorporating the three regularization techniques mentioned above. Thereafter we can maximize the log likelihood to compute the parameters of the K theme models.

However, before we estimate the parameter of the theme models, we refine our framework by regularizing the themes temporally as well as due to co-participation of participants. This is discussed in the following two sub-sections.

Temporal Regularization. We incorporate temporal characterization of themes in our theme model [27]. We conjecture that a word in the chunk can be attributed either to the textual context of the chunk $\lambda_{i,q}$, or the time slice q—for example, certain words can be highly popular on certain time slices due to related external events. Hence the theme associated with words in a chunk $\lambda_{i,q}$ needs to be regularized with respect to the time slice q. We represent the chunk $\lambda_{i,q}$ at time slice q with the probabilistic mixture model:

$$p(w : \lambda_{i,q}, q) = \sum_{j=1}^{K} p(w, \theta_j | \lambda_{i,q}, q) \tag{1}$$

where w is a word in the chunk $\lambda_{i,q}$ and θ_j is the j-th theme. The joint probability on the right hand side can be decomposed as:

$$\begin{aligned} p(w, \theta_j | \lambda_{i,q}, q) &= p(w|\theta_j) \cdot p(\theta_j | \lambda_{i,q}, q) \\ &= p(w|\theta_j) \cdot ((1-\gamma_q) \cdot p(\theta_j | \lambda_{i,q}) + \gamma_q \cdot p(\theta_j | q)), \end{aligned} \tag{2}$$

where γ_q is a parameter that regulates the probability of a theme θ_j given the chunk $\lambda_{i,q}$ and the probability of a theme θ_j given the time slice q. Note that since a conversation can alternatively be represented as a set of chunks, the collection of all chunks over all conversations is simply the set of conversations C. Hence the log likelihood of the entire collection of chunks is equivalent to the likelihood of the M conversations in C, given the theme model. Weighting the log likelihood of the model parameters with the occurrence of different words in a chunk, we get the following equation:

$$L(C) = \log p(C) = \sum_{\lambda_{i,q} \in C} \sum_{w \in \lambda_{i,q}} n(w, \lambda_{i,q}) \cdot \log \sum_{j=1}^{K} p(w, \theta_j | \lambda_{i,q}, q), \tag{3}$$

where $n(w, \lambda_{i,q})$ is the count of the word w in the chunk $\lambda_{i,q}$ and $p(w, \theta_j | \lambda_{i,q}, q)$ is given by eqn. 2. However, the theme distributions of two chunks of a conversation across two consecutive time slices should not too divergent from each other. That is, they need to be temporally smooth. For a particular topic θ_j this smoothness is thus based on minimization of the following L^2 distance between its probabilities across every two consecutive time slices:

$$d_T(j) = \sum_{q=2}^{Q} (p(\theta_j | q) - p(\theta_j | q-1))^2. \tag{4}$$

Incorporating this distance in eqn. 3 we get a new log likelihood function which smoothes all the K theme distributions across consecutive time slices:

$$L_1(C) = \sum_{\lambda_{i,q} \in C} \sum_{w \in \lambda_{i,q}} n(w, \lambda_{i,q}) \cdot \log \sum_{j=1}^{K} (p(w, \theta_j | \lambda_{i,q}, q) + \exp(-d_T(j))). \tag{5}$$

Now we discuss how this theme model is further regularized to incorporate prior knowledge about co-participation of individuals in the conversations.

Co-participation based Regularization. Our intuition behind this regularization is based on the idea that if several participants comment on a pair of chunks, then their theme distributions are likely to be closer to each other.

To recall, chunks being representations of conversations at a particular time slice, we therefore define a participant co-occurrence graph $G(C, E)$ where each vertex in C is a conversation c_i and an undirected edge $e_{i,m}$ exists between two conversations c_i and c_m if they share at least one common participant. The edges are also associated with weights $\omega_{i,m}$ which define the fraction of common participants between two conversations. We incorporate participant-based regularization based on this graph by minimizing the distance between the edge weights of two adjacent conversations with respect to their corresponding theme distributions.

The following regularization function ensures that the theme distribution functions of conversations are very close to each other if the edge between them in the participant co-occurrence graph G has a high weight:

$$R(C) = \sum_{c_i, c_m \in C} \sum_{j=1}^{K} (\omega_{i,m} - (1 - (f(\theta_j | c_i) - f(\theta_j | c_m))^2))^2, \tag{6}$$

where $f(\theta_j | c_i)$ is defined as a function of the theme θ_j given the conversation c_i and the L^2 distance between $f(\theta_j | c_i)$ and $f(\theta_j | c_m)$ ensures that the theme distributions of adjacent conversations are similar. Since a conversation is associated with multiple chunks, thus $f(\theta_j|c_i)$ is given as in [26]:

$$f(\theta_j|c_i) = p(\theta_j|c_i) = \sum_{\lambda_{i,q} \in c_i} p(\theta_j|\lambda_{i,q}) \cdot p(\lambda_{i,q}|c_i). \tag{7}$$

Now, using eqn. 5 and eqn. 6, we define the final combined optimization function which minimizes the negative of the log likelihood and also minimizes the distance between theme distributions with respect to the edge weights in the participant co-occurrence graph:

$$O(C) = -(1-\varsigma) \cdot L_1(C) + \varsigma \cdot R(C), \tag{8}$$

where the parameter ς controls the balance between the likelihood using the multinomial theme model and the smoothness of theme distributions over the participant graph. It is easy to note that when $\varsigma = 0$, then the objective function is the temporally regularized log likelihood as in eqn. 5. When $\varsigma = 1$, then the objective function yields themes which are smoothed over the participant co-occurrence graph. Minimizing $O(C)$ for $0 \leq \varsigma \leq 1$ would give us the theme models that best fit the collection.

Now to learn the hidden parameters of the theme model in eqn. 8, we use a different technique of parameter estimation based on the Generalized Expectation Maximization algorithm (GEM [26]). Details of the estimation can be referred to in [13].

3.3 Interestingness

In this section we describe our interestingness models and then discuss a method that jointly optimizes the two types of interestingness incorporating temporal smoothness.

3.3.1 Interestingness of Participants

We pose the problem of determining the interestingness of a participant at a certain time slice as a simple one-dimensional random walk model where she communicates either based on her past history of communication behavior in the previous time slice, or relies on her independent desire of preference over different themes (random jump). This formulation is described in Figure 3.

We conjecture that the state signifying the past history of communication behavior of a participant i at a certain time slice q, denoted as $\mathbf{A}(q-1)$ comprises the variables: (a) whether she was interesting in the previous time slice, $\mathbf{I_P}^{(q-1)}(i)$, (b) whether her comments in the past impacted other participants to communi-

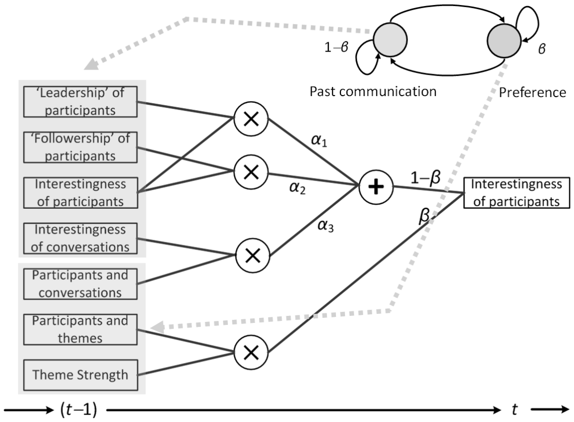

1−β
β
Past communication
Preference
'Leadership' of participants
'Followership' of participants
Interestingness of participants
Interestingness of conversations
Participants and conversations
Participants and themes
Theme Strength
α1
α2
α3
1−β
β
Interestingness of participants
(t–1)
t

Fig. 3 Random walk model for determining interestingness of participants.

cate and their interestingness measures, $\mathbf{P_F}^{(q-1)}(i,:)\cdot\mathbf{I_P}^{(q-1)2}$, (c) whether she followed several interesting people in conversations at the previous time slice $q-1$, $\mathbf{P_L}^{(q-1)}(i,:)\cdot\mathbf{I_P}^{(q-1)}$, and (d) whether the conversations in which she participated became interesting in the previous time slice $q-1$, $\mathbf{P_C}^{(q-1)}(i;:)\cdot\mathbf{I_C}^{(q1)}$. The independent desire of a participant i to communicate is dependent on her theme distribution and the strength of the themes at the previous time slice $q-1$: $\mathbf{P_T}^{(q-1)}(i,:)\cdot\mathbf{T_S}^{(q-1)}$.

Thus the recurrence relation for the random walk model to determine the interestingness of all participants at time slice q is given as:

$$\mathbf{I_P}^{(q)} = (1-\beta)\cdot\mathbf{A}^{(q-1)} + \beta\cdot(\mathbf{P_T}^{(q-1)}\cdot\mathbf{T_S}^{(q-1)}), \tag{9}$$

where,

$$\mathbf{A}^{(q-1)} = \alpha_1\cdot\mathbf{P_L}^{(q-1)}\cdot\mathbf{I_P}^{(q-1)} + \alpha_2\cdot\mathbf{P_F}^{(q-1)}\cdot\mathbf{I_P}^{(q-1)} + \alpha_3\cdot\mathbf{P_C}^{(q-1)}\cdot\mathbf{I_C}^{(q1)}. \tag{10}$$

Here α_1, α_2 and α_3 are weights that determine mutual relationship between the variables of the past history of communication state $\mathbf{A}^{(q-1)}$, and β the transition parameter of the random walk that balances the impact of past history and the random jump state involving participant's independent desire to communicate. In this paper, β is empirically set to be 0.5.

3.3.2 Interestingness of Conversations

Similar to interestingness of participants, we pose the problem of determining the interestingness of a conversation as a random walk where a conversation can be-

[2] To recall, $\mathbf{X}(i,:)$ is the i-th row of the 2-dimensional matrix $\mathbf{X}$.

come interesting based on two states as shown in Figure 4. Hence to determine the interestingness of a conversation i at time slice q, we conjecture that it depends on whether the participants in conversation i became interesting at $q-1$, given as, $\mathbf{P_C}^{(q-1)}(i,:)^t \cdot \mathbf{I_P}^{(q-1)}$, or whether the conversations belonging to the strong themes in $q-1$ became interesting, which is given as, $diag(\mathbf{C_T}^{(q-1)}(i,:) \cdot \mathbf{T_S}^{(q-1)}) \cdot \mathbf{I_C}^{(q-1)}$. Thus the recurrence relation of interestingness of all conversations at time slice q is:

$$\mathbf{I_C}^{(q)} = \psi \cdot \mathbf{P_C}^{(q-1)^t} \cdot \mathbf{I_P}^{(q-1)} + (1-\psi) \cdot diag(\mathbf{C_T}^{(q-1)} \cdot \mathbf{T_S}^{(q-1)}) \cdot \mathbf{I_C}^{(q-1)}, \quad (11)$$

where ψ is the transition parameter of the random walk that balances the impact of interestingness due to participants and due to themes. Clearly, when $\psi = 1$, the interestingness of conversation depends solely on the interestingness of the participants at $q-1$; and when $\psi = 1$, the interestingness depends on the theme strengths in the previous time slice $q-1$.

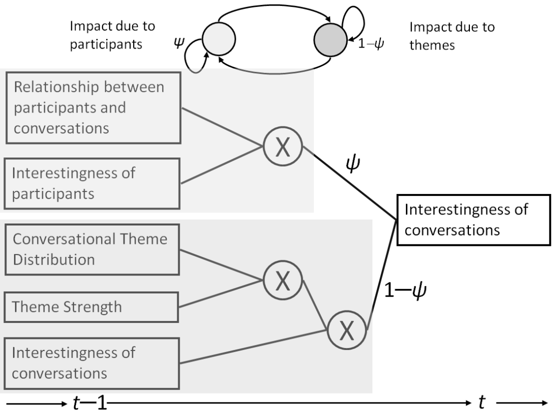

Impact due to participants
ψ
$1-\psi$
Impact due to themes
Relationship between participants and conversations
Interestingness of participants
X
ψ
Interestingness of conversations
Conversational Theme Distribution
Theme Strength
Interestingness of conversations
X
X
$1-\psi$
$t-1$
t

Fig. 4 Random walk model for determining interestingness of conversations.

3.3.3 Joint Optimization of Interestingness

We observe that the measures of interestingness of participants and of conversations described in previous sections involve several free (unknown) parameters. In order to determine optimal values of interestingness, we need to learn the weights α_1, α_2 and α_3 in eqn. 10 and the transition probability for the conversations in eqn. 11. Moreover, the optimal measures of interestingness should ensure that the variations in their values are smooth over time. Hence we present a novel joint optimization framework, which maximizes the two interestingness measures for optimal $(\alpha_1, \alpha_2, \alpha_3, \psi)$ and also incorporates temporal smoothness.

The joint optimization framework is based on the idea that the optimal parameters in the two interestingness equations are those which maximize the interestingness of participants and of conversations jointly. Let us denote the set of the parameters to be optimized as the vector, $\mathbf{X} = [\alpha_1, \alpha_2, \alpha_3, \psi]$. We can therefore represent $\mathbf{I_P}$ and $\mathbf{I_C}$ as functions of $\mathbf{X}$. We define the following objective function $g(\mathbf{X})$ to estimate $\mathbf{X}$ by maximizing $g(\mathbf{X})$:

$$g(\mathbf{X}) = \rho \cdot \|\mathbf{I_P}(\mathbf{X})\|^2 + (1-\rho) \cdot \|\mathbf{I_C}(\mathbf{X})\|^2, \tag{12}$$

s.t. $0 \leq \psi 1, \alpha_1, \alpha_2, \alpha_3 \geq 0, \mathbf{I_P} \geq 0, \mathbf{I_C} \geq 0, \alpha_1 + \alpha_2 + \alpha_3 = 1$.

In the above function, ρ is an empirically set parameter to balance the impact of each interestingness measure in the joint optimization. Now to incorporate temporal smoothness of interestingness in the above objective function, we define a L^2 norm distance between the two interestingness measures across all consecutive time slices q and $q-1$:

$$\begin{aligned} d_P &= \sum_{q=2}^{Q} (\|\mathbf{I_P}^{(q)}(\mathbf{X})\|^2 - \|\mathbf{I_P}^{(q-1)}(\mathbf{X})\|^2), \\ d_C &= \sum_{q=2}^{Q} (\|\mathbf{I_C}^{(q)}(\mathbf{X})\|^2 - \|\mathbf{I_C}^{(q-1)}(\mathbf{X})\|^2). \end{aligned} \tag{13}$$

We need to minimize these two distance functions to incorporate temporal smoothness. Hence we modify our objective function,

$$g_1(\mathbf{X}) = \rho \cdot \|\mathbf{I_P}(\mathbf{X})\|^2 + (1-\rho) \cdot \|\mathbf{I_C}(\mathbf{X})\|^2 + \exp(-d_P) + \exp(d_C), \tag{14}$$

where $0 \leq \psi 1, \alpha_1, \alpha_2, \alpha_3 \geq 0, \mathbf{I_P} \geq 0, \mathbf{I_C} \geq 0, \alpha_1 + \alpha_2 + \alpha_3 = 1$.

Maximizing the above function $g_1(\mathbf{X})$ for optimal $\mathbf{X}$ is equivalent to minimizing $-g_1(\mathbf{X})$. Thus this minimization problem can be reduced to a convex optimization form because (a) the inequality constraint functions are also convex, and (b) the equality constraint is affine. The convergence of this optimization function is skipped due to space limit.

Now, the minimum value of $-g_1(\mathbf{X})$ corresponds to an optimal $\mathbf{X}$* and hence we can easily compute the optimal interestingness measures $\mathbf{I_P}$* and $\mathbf{I_C}$* for the optimal $\mathbf{X}$*. Given our framework for determining interestingness of conversations, we now discuss the measures of consequence of interestingness followed by extensive experimental results.

3.4 Consequences of Interestingness

An interesting conversation is likely to have consequences. These include the (commenting) activity of the participants, their cohesiveness in communication and an effect on the interestingness of the themes. It is important to note here that the consequence is generally felt at a future point of time; that is, it is associated with a certain time lag (say, δ days) with respect to the time slice a conversation becomes interesting (say, q). Hence we ask the following three questions related to the future consequences of an interesting conversation:

Activity. Do the participants in an interesting conversation i at time q take part in other conversations relating to similar themes at a future time, $q+\delta$ We define this as follows,

$$Act^{q+\delta}(i) = \frac{1}{\varphi_{i,q+\delta}} \sum_{k=1}^{|\varphi_{i,q+\delta}|} \sum_{j=1}^{|P_{i,q}|} \mathbf{P_C}^{(q+\delta}(j,k), \tag{15}$$

where $P_{i,q}$ is the set of participants on conversation i at time slice q, and $\varphi_{i,q+\delta}$ is the set of conversations m such that, $m \in \varphi_{i,q+\delta}$ if the KL-divergence of the theme distribution of m at time $q+\delta$ from that of i at q is less than an empirically set threshold: $D(C_T^{(q)}(i,:)||C_T^{(q+\delta)}(m,:)) \leq \varepsilon$.

Cohesiveness. Do the participants in an interesting conversation i at time q exhibit cohesiveness in communication (co-participate) in other conversations at a future time slice, $q+\delta$ In order to define cohesiveness, we first define co-participation of two participants, j and k as,

$$O^{(q+\delta)}(j;k) = \frac{\mathbf{P_P}^{(q+\delta)}(j,k)}{\mathbf{P_C}^{(q+\delta)}(j,k)}, \tag{16}$$

where $\mathbf{P_P}^{(q+\delta)}(j,k)$ is defined as the participant-participant matrix of co-participation constructed as, $P_C^{(q+\delta)} \cdot (P_C^{(q+\delta)})^t$. Hence the cohesiveness in communication at time $q+\delta$ between participants in a conversation i is defined as,

$$Co^{(q+\delta)}(i) = \frac{1}{|P_{i,q}|} \sum_{j=1}^{P_{i,q}} \sum_{k=1}^{|P_{i,q}|} O^{(q+\delta)}(j;k). \tag{17}$$

Thematic Interestingness. Do other conversations having similar theme distribution as the interesting conversation c_i (at time q), also become interesting at a future time slice $q+\delta$ We define this consequence as thematic interestingness and it is given by,

Table 2 Political events in the time period of analysis.

DATE	EVENT
Jul 23'08	Obama makes trip to the Europe and Middle East
Aug 29'08	Alaska Governor Sarah Palin is selected by McCain as his choice for the Republican VP candidate
Sep 1'08	2008 Republican National Convention convenes in Minneapolis-St.Paul, Minnesota
Sep 15'08	Lehman Brothers goes bankrupt, Merrill Lynch is dissolved
Sep 24'08	President Bush addresses the nation on the financial crisis

$$TInt^{(q+delta)}(i) = \frac{1}{\varphi_{i,q+delta}} \sum_{j=1}^{|\varphi_{i,q+delta}|} I_C^{(q+\delta)}(j). \qquad (18)$$

To summarize, we have developed a method to characterize interestingness of conversations based on the themes, and the interestingness property of the participants. We have jointly optimized the two types of interestingness to get optimal interestingness of conversations. And finally we have discussed three metrics which account for the consequential impact of interesting conversations. Now we would discuss the experimental results on this model.

3.5 Experimental Studies

The experiments performed to test our model are based on a dataset from the largest video-sharing site, YouTube, which serves as a rich source of online conversations associated with shared media elements. We crawled a total set of 132,348 videos involving 8,867,284 unique participants and 89,026,652 comments over a period of 15 weeks from June 20, 2008 to September 26, 2008. Now we discuss the results of experiments conducted to test our framework. First we present the results on the interestingness of participants, followed by that of conversations.

The results of interestingness of the participants of conversations are shown in a visualization in Figure 5. We have visualized a set of 45 participants over the period of 15 weeks by pooling the top three most interesting participants from each week. The participants are shown column-wise in the visualization with decreasing mean number of comments written from left to right. The intensity of the red block represents the degree of interestingness of a participant at a particular time slice. The figure also shows plots of the comment distribution and the interestingness distributions for the participants at each time slice.

In order to analyze the dynamics of interestingness, we also qualitatively observe its association with a set of external events collected from The New York Times, related to Politics. The events along with their dates are shown in Table 2.

From the results of interestingness of participants, we observe that interestingness closely follows the number of comments on weeks which are not associated with significant external events (weeks 1-4, 6-10). Whereas on other weeks, especially the last three weeks 13, 14 and 15, we observe that there are several politi-

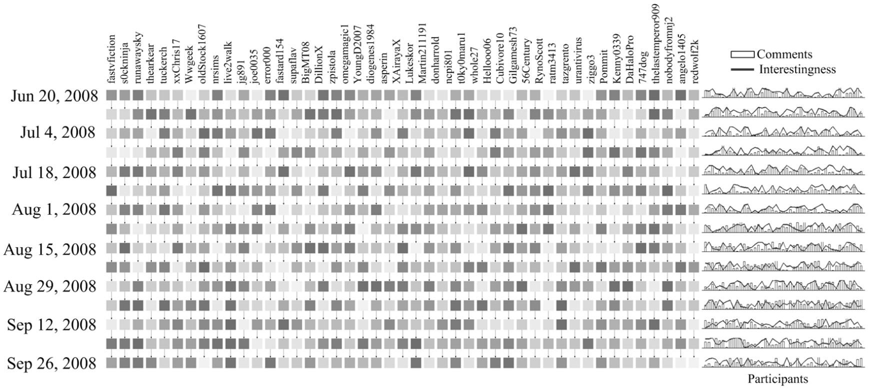


Fig. 5 Interestingness of 45 participants from YouTube, ordered by decreasing number of comments from left to right, is visualized. Interestingness is less affected by number of comments during periods of several external events..

cal happenings and as a result the interestingness distribution of participants does not seem to follow well the comment distribution. Hence we conclude that during periods of significant external events, participants can become interesting despite writing fewer comments—high interestingness can instead be explained due to their preference for the conversational theme which reflects the external event.

The results of the dynamics of interestingness of conversations are shown in Figure 6. We conceive a similar visualization as Figure 5 presented previously. Conversations are shown column-wise and time row-wise (15 weeks). A set of 45 conversations are pooled based on the top three most interesting conversations at each week. From left to right, the conversations are shown with respect to decreasing number of comments. We also show a temporal plot of the mean interestingness per week in order to understand the relationship of interestingness to external happening from Table 2.

From the visualization in Figure 6, we observe that the mean interestingness of conversations increase significantly during weeks 11-15. This is explained when we observe the association with large number of political happening in the said period (Table 2). Hence we conclude that conversations in general become more interesting when there are significant events in the external world—an artifact that online conversations are reflective of chatter about external happenings.

In closing for this problem, note that today there is significant online chatter, discussion and thoughts that are expressed over shared rich media artifacts, e.g. photos, videos etc, often reflecting public sentiment on socio-political events. While different media sites can provide coverage over the same information content with variable degrees of associated chatter, it becomes imperative to determine suitable methods and techniques to identify which media sources are likely to provide information that can be deemed to be "interesting" to a certain user. Suppose a user Alice is interested in identifying "interesting" media sources dissipating information on public sentiments regarding the recent elections in Iran back in 2009. To

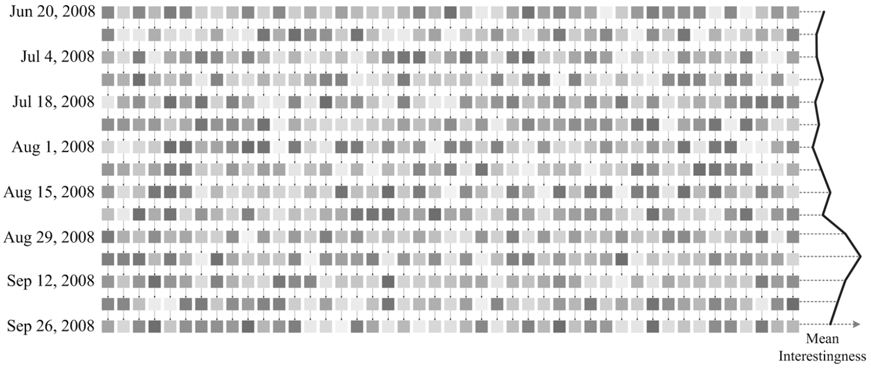


Fig. 6 Interestingness of 45 conversations from YouTube, ordered by decreasing number of comments from left to right, is visualized. Mean interestingness of conversations increases during periods of several external events.

serve Alice's needs, we need to be able to characterize chatter or conversations that emerge centered around rich media artifacts, that she would find useful. We believe the proposed framework can serve the needful to tackle the modern day information needs on the social Web.

Nevertheless, it goes without saying that human communication activity, manifested via such "conversations" involves mutual exchange of information, and the pretext of any social interaction among a set of individuals is a reflection of how our behavior, actions and knowledge can be modified, refined, shared or amplified based on the information that flows from one individual to another. Thus, over several decades, the structure of social groups, society in general and the relationships among individuals in these societies have been shaped to a great extent by the *flow of information* in them. Diffusion is hence the process by which a piece of information, an idea or an innovation flows through certain communication channels over time among the individuals in a social system.

The pervasive use of online social media has made the cost involved in propagating a piece of information to a large audience extremely negligible, providing extensive evidences of large-scale social contagion. There are multifaceted personal publishing modalities available to users today, where such large scale social contagion is prevalent: such as weblogs, social networking sites like MySpace and Facebook as well as microblogging tools such as Twitter. These communication tools are open to frequent widespread observation to millions of users, and thus offer an inexpensive opportunity to capture large volumes of information flows at the individual level. If we want to understand the extent to which ideas are adopted via these communication affordances provided by different online social platforms, it is important to understand the extent to which people are likely to be affected by decisions of their friends and colleagues, or the extent to which "word-of-mouth" effects will take hold via communication. In the following section we propose models of diffusion of information in the light of how similar user attributes, that embody observed "homophily" in networks, affect the overall social process.

4 Information Diffusion

The central goal in this section is to investigate the relationship between homophily among users and the social process of information diffusion. By "homophily," we refer to the idea that users in a social system tend to bond more with ones who are "similar" to them than ones who are dissimilar. The homophily principle has been extensively researched in the social sciences over the past few decades [7, 25, 24]. These studies were predominantly ethnographic and cross-sectional in nature and have revealed that homophily *structures* networks. That is, a person's ego-centric social network is often homogeneous with regard to diverse social, demographic, behavioral, and intra-personal characteristics [24] or revolves around social foci such as co-location or commonly situated activities [14]. Consequently, in the context of physical networks, these works provide evidence that the existence of homophily is likely to impact the information individuals receive and propagate, the communication activities they engage in, and the social roles they form.

Homophilous relationships have also been observed on online media such as Facebook, Twitter, Digg and YouTube. These networks facilitate the sharing and propagation of information among members of their networks. In these networks, homophilous associations can have a significant impact on *very large scale* social phenomena, including group evolution and information diffusion. For example, the popular social networking site Facebook allows users to engage in community activities via homophilous relationships involving common organizational affiliations. Whereas on the fast-growing social media Twitter, several topics such as '#Elections2008', '#MichaelJackson', 'Global Warming' etc have historically featured extensive postings (also known as "tweets") due to the common interests of large sets of users in politics, music and environmental issues respectively.

These networks, while diverse in terms of their affordances (i.e. what they allow users to do), share some common features. First, there exists a social action (e.g. posting a tweet on Twitter) within a shared social space (i.e. the action can be observed by all members of the users' contact network), that facilitates a social process (e.g. diffusion of information). Second, these networks expose attributes including location, time of activity and gender to other users. Finally, these networks also reveal these users attributes as well as the communication, to third party users (via the API tools); thus allowing us to study the impact of a specific attribute on information diffusion within these networks.

The study of the impact of homophily on information diffusion can be valuable in several contexts. Today, due to the plethora of diverse retail products available online to customers, advertising is moving from the traditional "word-of-mouth" model, to models that exploit interactions among individuals on social networks. To this effect, previously, some studies have provided useful insights that social relationships impact the adoption of innovations and products [19]. Moreover there has been theoretical and empirical evidence in prior work [36] that indicates that individuals have been able to transmit information through a network (via messages) in a sufficiently small number of steps, due to homophily along recognizable personal identities. Hence a viral marketer attempting to advertise a new product could benefit from considering specific sets of users on a social space who are *homophilous* with respect to their interest in similar products or features. Other contexts in which understanding the role of homophily in information diffusion can be important, include, disaster mitigation during crisis situations, understanding social roles of users and in leveraging distributed social search.

4.1 Preliminaries

4.1.1 Social Graph Model

We define our social graph model as a directed graph $G(V,E)^3$, such that V is the set of users and $e_{ij} \in E$ if and only if user u_i and u_j are "friends" of each other (bi-directional contacts). Let us further suppose that each user $u_i \in V$ can perform a set of "social actions", $\mathscr{O} = \{O_1, O_2, \ldots\}$, e.g. posting a tweet, uploading a photo on Flickr or writing on somebody's Facebook Wall. Let the users in V also be associated with a set of attributes $\mathscr{A} = \{a_k\}$ (e.g. location or organizational affiliation) that are responsible for homophily. Corresponding to each value υ defined over an attribute $a_k \in \mathscr{A}$, we construct a social graph $G(a_k = \upsilon)$ such that it consists of the users in G with the particular value of the attribute, while an edge exists between two users in $G(a_k)$ if there is an edge between them in G.[4] E.g., for location, we can define sets of social graphs over users from Europe, Asia etc.

In this section, our social graph model is based on the social media Twitter. Twitter features a micro-blogging service that allows users to post short content, known as "tweets", often comprising URLs usually encoded via bit.ly, tinyurl, etc. The particular "social action" in this context is the posting of a tweet; also popularly called "tweeting". Users can also "follow" other users; hence if user u_i follows u_j, Twitter allows u_i to subscribe to the tweets of u_j via feeds; u_i is then also called a "follower" of u_j. Two users are denoted as "friends" on Twitter if they "follow" each other. Note that, in the context of Twitter, using the bi-directional "friend" link is more useful compared to the uni-directional "follow" link because the former is more likely to be robust to spam—a normal user is less likely to follow a spam-like account. Further, for the particular dataset of Twitter, we have considered a set of four attributes associated with the users:

[3] Henceforth referred to as the baseline social graph G.

[4] For simplicity, we omit specifying the attribute value υ in the rest of the section, and refer to $G(a_k = \upsilon)$ as the "attribute social graph" $G(a_k)$.

Location of users, extracted using the timezone attribute of Twitter users. Specifically, the values of location correspond to the different continents, e.g. Asia, Europe and North America.

Information roles of users, we consider three categories of roles: "generators", "mediators" and "receptors". Generators are users who create several posts (or tweets) but few users respond to them (via the @ tag on Twitter, which is typically used with the username to respond to a particular user, e.g. @BillGates). While receptors are those who create fewer posts but receive several posts as responses. Mediators are users who lie between these two categories.

Content creation of users, we use the two content creation roles: "meformer" (users who primarily post content relating to self) and "informer" (users posting content about external happenings) as discussed in [29].

Activity behavior of users, i.e. the distribution of a particular social action over a certain time period. We consider the mean number of posts (tweets) per user over 24 hours and compute similarities between pairs of users based on the Kullback-Leibler (KL) divergence measure of comparing across distributions.

4.1.2 Attribute Homophily

Attribute homophily [25, 24] is defined as the tendency of users in a social graph to associate and bond with others who are "similar" to them along a certain attribute or contextual dimension e.g. age, gender, race, political view or organizational affiliation. Specifically, a pair of users can be said to be "homophilous" if one of their attributes match in a proportion greater than that in the network of which they are a part. Hence in our context, for a particular value of $a_k \in \mathscr{A}$, the users in the social graph $G(a_k)$ corresponding to that value are homophilous to each other.

4.1.3 Topic Diffusion

Diffusion with respect to a particular topic at a certain time is given as the flow of information on the topic from one user to another via the social graph, and based on a particular social action. Specifically,

Definition 1. Given two users u_i and u_j in the baseline social graph G such that $e_{ij} \in E$, there is diffusion of information on topic θ from u_j to u_i if u_j performs a particular social action O_r related to θ at a time slice t_{m-1} and is succeeded by u_i in performing the same action on θ at the next time slice t_m, where $t_{m-1} < t_m$.[5]

[5] Since we discuss our problem formulation and methodology for a specific social action, the dependence of different concepts on O_r is omitted in the rest of the section for simplicity.

Further, topic diffusion subject to homophily along the attribute a_k is defined as the diffusion over the attribute social graph $G(a_k)$.

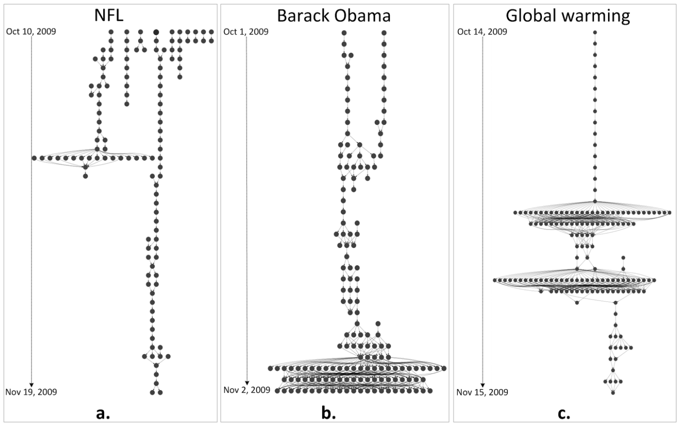


Fig. 7 Example of different diffusion series from Twitter on three different topics. The nodes are users involved in diffusion while the edges represent "friend links" connecting two users.

In the context of Twitter, topic diffusion can manifest itself through three types of evidences: (1) users posting tweets using the same URL, (2) users tweeting with the same hashtag (e.g. #MichaelJackson) or a set of common keywords, and (3) users using the re-tweet (RT) symbol. We utilize all these three cases of topic diffusion in this work.

4.1.4 Diffusion Series

In order to characterize diffusion, we now define a topology called a *diffusion series*[6] that summarizes diffusion in a social graph for a given topic over a period of time. Formally,

Definition 2. A diffusion series $s_N(\theta)$ on topic θ and over time slices t_1 to t_N is defined as a directed acyclic graph where the nodes represent a subset of users in the baseline social graph G, who are involved in a specific social action O_r over θ at any time slice between t_1 and t_N.

Note, in a diffusion series $s_N(\theta)$ a node represents an occurrence of a user u_i creating at least one instance of the social action O_r about θ at a certain time slice t_m such that $t_1 \leq t_m \leq t_N$. Nodes are organized into "slots"; where nodes associated with the same time slice t_m are arranged into the same slot l_m. Hence it is possible

[6] Note, a diffusion series is similar to a diffusion tree as in [23, 4], however we call it a "series" since it is constructed progressively over a period of time and allows a node to have multiple sources of diffusion.

that the same user is present at multiple slots in the series if s/he tweets about the same topic θ at different time slices. Additionally, there are edges between nodes across two adjacent slots, indicating that user u_i in slot l_m performs the social action O_r on θ at t_m, after her friend u_j has performed action on the same topic θ at the previous time slice t_{m-1} (i.e. at slot l_{m-1}). There are no edges between nodes at the same slot l_m: a diffusion series $s_N(\theta)$ in this work captures diffusion on topic θ *across* time slices, and does not include possible flow occurring at the same time slice.

For the Twitter dataset, we have chosen the granularity of the time slice t_m to be sufficiently small, i.e. a day to capture the dynamics of diffusion. Thus all the users at slot l_m tweet about θ on the same day; and two consecutive slots have a time difference of one day. Examples of different diffusion series constructed on topics from Twitter have been shown in Figure 7.

Since each topic θ can have multiple disconnected diffusion series $s_N(\theta)$ at any given time slice t_N, we call the family of all diffusion series a *diffusion collection* $\mathscr{S}_N(\theta) = \{s_N(\theta)\}$. Corresponding to each value of the attribute a_k, the diffusion collection over the attribute social graph $G(a_k)$ at t_N is similarly given as $\mathscr{S}_{N;a_k}(\theta) = \{s_{N;a_k}(\theta)\}$.

4.2 Problem Statement

Given, (1) a baseline social graph $G(V,E)$; (2) a set of social actions $\mathscr{O} = \{O_1, O_2, \ldots\}$ that can be performed by users in V, and (3) a set of attributes $\mathscr{A} = \{a_k\}$ that are shared by users in V, we perform the following two preliminary steps. First, we construct the attribute social graphs $\{G(a_k)\}$, for all values of $a_k \in \mathscr{A}$. Second, we construct diffusion collections corresponding to G and $\{G(a_k)\}$ for a given topic θ (on which diffusion is to be estimated over time slices t_1 to t_N) and a particular social action O_r: these are given as $\mathscr{S}_N(\theta)$ and $\{\mathscr{S}_{N;a_k}(\theta)\}$ respectively. The technical problem addressed in this section involves the following:

1. *Characterization:* Based on each of the diffusion collections $\mathscr{S}_N(\theta)$ and $\{\mathscr{S}_{N;a_k}(\theta)\}$, we extract diffusion characteristics on θ at time slice t_N given as: $\mathbf{d}_N(\theta)$ and $\{\mathbf{d}_{N;a_k}(\theta)\}$ respectively (section 4.3);
2. *Prediction:* We predict the set of users likely to perform the same social action at the next time slice t_{N+1} corresponding to each of the diffusion collections $\mathscr{S}_N(\theta)$ and $\{\mathscr{S}_{N;a_k}(\theta)\}$. This gives the diffusion collections at t_{N+1}: $\hat{\mathscr{S}}_{N+1}(\theta)$ and $\{\hat{\mathscr{S}}_{N+1;a_k}(\theta)\} \forall a_k \in \mathscr{A}$ (section 4.4);
3. *Distortion Measurement:* We extract diffusion characteristics at t_{N+1} over the (predicted) diffusion collections, $\hat{\mathscr{S}}_{N+1}(\theta)$ and $\{\hat{\mathscr{S}}_{N+1;a_k}(\theta)\}$, given as, $\hat{\mathbf{d}}_{N+1}(\theta)$ and $\{\hat{\mathbf{d}}_{N+1;a_k}(\theta)\}$ respectively. Now we quantify the impact of attribute homophily on diffusion based on two kinds of distortion measurements on $\hat{\mathbf{d}}_{N+1}(\theta)$ and $\{\hat{\mathbf{d}}_{N+1;a_k}(\theta)\}$. A particular attribute $a_k \in \mathscr{A}$ would have an impact on diffusion if $\hat{\mathbf{d}}_{N+1;a_k}(\theta)$, avergaed over all possible values of a_k: (a) has lower distortion

with respect to the actual (i.e. $\mathbf{d}_{N+1}(\theta)$); and (b) can quantify external time series (search, news trends) better, compared to either $\hat{\mathbf{d}}_{N+1}(\theta)$ or $\{\hat{\mathbf{d}}_{N+1;a'_k}(\theta)\}$, where $k' \neq k$ (section 4.7).

4.3 Characterizing Diffusion

We describe eight different measures for quantifying diffusion characteristics given by the baseline and the attribute social graphs on a certain topic and via a particular social action.

Volume: Volume is a notion of the overall degree of contagion in the social graph. For the diffusion collection $\mathscr{S}_N(\theta)$ over the baseline social graph G, we formally define volume $v_N(\theta)$ with respect to θ and at time slice t_N as the ratio of $n_N(\theta)$ to $\eta_N(\theta)$, where $n_N(\theta)$ is the total number of users (nodes) in the diffusion collection $\mathscr{S}_N(\theta)$, and $\eta_N(\theta)$ is the number of users in the social graph G associated with θ.

Participation: Participation $p_N(\theta)$ at time slice t_N [4] is the ratio of the number of non-leaf nodes in the diffusion collection $\mathscr{S}_N(\theta)$, normalized by $\eta_N(\theta)$.

Dissemination: Dissemination $\delta_N(\theta)$ at time slice t_N is given by the ratio of the number of users in the diffusion collection $\mathscr{S}_N(\theta)$ who do not have a parent node, normalized by $\eta_N(\theta)$. In other words, they are the "seed users" or ones who get involved in the diffusion due to some unobservable external influence, e.g. a news event.

Reach: Reach $r_N(\theta)$ at time slice t_N [23] is defined as the ratio of the mean of the number of slots to the sum of the number of slots in all diffusion series belonging to $\mathscr{S}_N(\theta)$.

Spread: For the diffusion collection $\mathscr{S}_N(\theta)$, spread $s_N(\theta)$ at time slice t_N [23] is defined as the ratio of the maximum number of nodes at any slot in $s_N(\theta) \in \mathscr{S}_N(\theta)$ to $n_N(\theta)$.

Cascade Instances: Cascade instances $c_N(\theta)$ at time slice t_N is defined as the ratio of the number of slots in the diffusion series $s_N(\theta) \in \mathscr{S}_N(\theta)$ where the number of *new* users at a slot l_m (i.e. non-occurring at a previous slot) is greater than that at the previous slot l_{m-1}, to $L_N(\theta)$, the number of slots in $s_N(\theta) \in \mathscr{S}_N(\theta)$.

Collection Size: Collection size $\alpha_N(\theta)$ at time slice t_N is the ratio of the number of diffusion series $s_N(\theta)$ in $\mathscr{S}_N(\theta)$ over topic θ, to the total number of connected components in the social graph G.

Rate: We define rate $\gamma_N(\theta)$ at time slice t_N as the "speed" at which information on θ diffuses in the collection $\mathscr{S}_N(\theta)$. It depends on the difference between the median time of posting of tweets at all consecutive slots l_m and l_{m-1} in the diffusion series $s_N(\theta) \in \mathscr{S}_N(\theta)$. Hence it is given as:

$$\gamma_N(\theta) = 1/(1+\frac{1}{L_N(\theta)}\sum_{l_{m-1},l_m\in\mathscr{S}_N(\theta)}(\bar{t}_m(\theta)-\bar{t}_{m-1}(\theta)), \tag{19}$$

where $\bar{t}_m(\theta)$ and $\bar{t}_{m-1}(\theta)$ are measured in seconds and $\bar{t}_m(\theta)$ corresponds to the median time of tweet at slot l_m in $s_N(\theta) \in \mathscr{S}_N(\theta)$.

These diffusion measures thus characterize diffusion at time slice t_N over $\mathscr{S}_N(\theta)$ as the vector: $\mathbf{d}_N(\theta) = [v_N(\theta), p_N(\theta), \delta_N(\theta), r_N(\theta), s_N(\theta), c_N(\theta), \alpha_N(\theta), \gamma_N(\theta)]$. Similarly, we compute the diffusion measures vector over $\{\mathscr{S}_{N;a_k}(\theta)\}$, given by: $\{\mathbf{d}_{N;a_k}(\theta)\}$, corresponding to each value of a_k.

4.4 Prediction Framework

In this section we present our method of predicting the users who would be part of the diffusion collections at a future time slice for the baseline and attribute social graphs. Our method comprises the following steps. (1) Given the observed diffusion collections until time slice t_N (i.e. $\mathscr{S}_N(\theta)$ and $\mathscr{S}_{N;a_k}(\theta)$), we first propose a probabilistic framework based on Dynamic Bayesian networks [30] to predict the users likely to perform the social action O_r at the next time slice t_{N+1}. This would yield us users at slot l_{N+1} in the different diffusion series at t_{N+1}. (2) Next, these predicted users give the diffusion collections at t_{N+1}: $\hat{\mathscr{S}}_{N+1}(\theta)$ and $\{\hat{\mathscr{S}}_{N+1;a_k}(\theta)\}$.

We present a Dynamic Bayesian network (DBN) representation of a particular social action by a user over time, that helps us predict the set of users likely to perform the social action at a future time (Figure 8(a)). Specifically, at any time slice t_N, a given topic θ and a given social action, the DBN captures the relationship between three nodes:

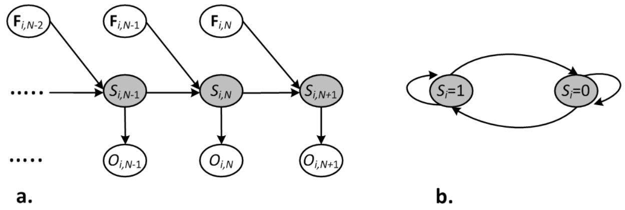


Fig. 8 (a) Structure of the Dynamic Bayesian network used for modeling social action of a user u_i. The diagram shows the relationship between environmental features ($\mathbf{F}_{i,N}(\theta)$), hidden states ($S_{i,N}(\theta)$) and the observed action ($O_{i,N}(\theta)$). (b) State transition diagram showing the 'vulnerable' ($S_i = 1$) and 'indifferent' states ($S_i = 0$) of a user u_i.

Environmental Features. That is, the set of contextual variables that effect a user u_i's decision to perform the action on θ at a future time slice t_{N+1} (given by $\mathbf{F}_{i,N}(\theta)$). It comprises three different measures: (1) u_i's degree of activity on θ in the past, given as the ratio of the number of posts (or tweets) by u_i on θ, to the total number of posts between t_1 and t_N; (2) mean degree of activity of u_i's friends in the past, given as the ratio of the number of posts by u_i's friends on θ, to the total number of posts by them between t_1 and t_N; and (3) popularity of topic θ at the previous time slice t_N, given as the ratio of the number of posts by all users on θ, to the total number of posts at t_N.

States. That is, latent states ($S_{i,N}(\theta)$) of the user u_i responsible for her involvement in diffusion at t_{N+1}. Our motivation in conceiving the latent states comes from the observation that, in the context of Twitter, a user can tweet on a topic under two kinds of circumstances: first, when she observes her friend doing so already: making her *vulnerable* to diffusion; and second, when her tweeting is *indifferent* to the activities of her friends. Hence the state node at t_{N+1} that impacts u_i's action can have two values as the vulnerable and the indifferent state (Figure 8(b)).

Observed Action. That is, evidence ($O_{i,N}(\theta)$) of the user u_i performing (or not performing) the action, corresponding values being: $\{1,0\}$ respectively.

Now we show how to predict the probability of the observed action at t_{N+1} (i.e. $\hat{O}_{i,N+1}(\theta)$) using $\mathbf{F}_{i,N}(\theta)$ and $S_{i,N+1}(\theta)$, based on the DBN model. Our goal is to estimate the following expectation[7]:

$$\hat{O}_{i,N+1} = E(O_{i,N+1}|O_{i,N},\mathbf{F}_{i,N}). \tag{20}$$

This involves computing $P(O_{i,N+1}|O_{i,N},\mathbf{F}_{i,N})$. This conditional probability can be written as an inference equation using the temporal dependencies given by the DBN and assuming first order Markov property:

$$\begin{aligned}
&P(O_{i,N+1}|O_{i,N},\mathbf{F}_{i,N}) \\
&= \sum_{S_{i,N+1}} \left[P(O_{i,N+1}|S_{i,N+1},O_{i,N},\mathbf{F}_{i,N}).P(S_{i,N+1}|O_{i,N},\mathbf{F}_{i,N})\right]. \\
&= \sum_{S_{i,N+1}} P(O_{i,N+1}|S_{i,N+1}).P(S_{i,N+1}|S_{i,N},\mathbf{F}_{i,N}).
\end{aligned} \tag{21}$$

Our prediction task thus involves two parts: predicting the probability of the hidden states given the environmental features, $P(S_{i,N+1}|S_{i,N},\mathbf{F}_{i,N})$; and predicting the probability density of the observation nodes given the hidden states, $P(O_{i,N+1}|S_{i,N+1})$, and thereby the expected value of observation nodes $\hat{O}_{i,N+1}$. These two steps are discussed in the following subsections.

[7] Without loss of generalization, we omit the topic θ in the variables in this subsection for the sake of simplicity.

4.5 Predicting Hidden States

Using Bayes rule, we apply conditional independence between the hidden states and the environmental features at the same time slice (ref. Figure 8(a)). The probability of the hidden states at t_{N+1} given the environmental features at t_N, i.e. $P(S_{i,N+1}|S_{i,N},\mathbf{F}_{i,N})$ can be written as:

$$P(S_{i,N+1}|S_{i,N},\mathbf{F}_{i,N}) \propto P(\mathbf{F}_{i,N}|S_{i,N}).P(S_{i,N+1}|S_{i,N}). \tag{22}$$

Now, to estimate the probability density of $P(S_{i,N+1}|S_{i,N},\mathbf{F}_{i,N})$ using eqn. 22 we assume that the hidden states $S_{i,N+1}$ follows a multinomial distribution over the environmental features $\mathbf{F}_{i,N}$ with parameter $\phi_{i,N}$, and a conjugate Dirichlet prior over the previous state $S_{i,N}$ with parameter $\lambda_{i,N+1}$. The optimal parameters of the pdf of $P(S_{i,N+1}|S_{i,N},\mathbf{F}_{i,N})$ can now be estimated using MAP:

$$\begin{aligned}
&\mathscr{L}(P(S_{i,N+1}|S_{i,N},\mathbf{F}_{i,N}))\\
&= \log(P(\mathbf{F}_{i,N}|S_{i,N})) + \log(P(S_{i,N+1}|S_{i,N}))\\
&= \log \mathbf{multinom}(\mathrm{vec}(\mathbf{F}_{i,N});\phi_{i,N})\\
&+ \log \mathbf{Dirichlet}(\mathrm{vec}(S_{i,N+1});\lambda_{i,N+1})\\
&= \log \frac{\sum_{jk}\mathbf{F}_{i,N;jk}!}{\prod_{jk}\mathbf{F}_{i,N;jk}!}\prod_{jk}\phi_{i,N;jk}^{\mathbf{F}_{i,N;jk}} + \log\frac{1}{B(\lambda_{i,N+1})}\prod_{jl}S_{i,N+1}^{S_{i,N;jl}}\\
&= \sum_{jk}\mathbf{F}_{i,N;jk}.\log\phi_{i,N;jk} + \sum_{jl}S_{i,N;jl}.\log S_{i,N+1;jl} + \text{const.}
\end{aligned} \tag{23}$$

where $B(\lambda_{i,N+1})$ is a beta-function with the parameter $\lambda_{i,N+1}$. Maximizing the log likelihood in eqn 23 hence yields the optimal parameters for the pdf of $P(S_{i,N+1}|S_{i,N},\mathbf{F}_{i,N})$.

4.6 Predicting Observed Action

To estimate the probability density of the observation nodes given the hidden states, i.e. $P(O_{i,N+1}|S_{i,N+1})$ we adopt a generative model approach and train two discriminative Hidden Markov Models—one corresponding to the class when u_i performs the action, and the other when she does not. Based on observed actions from t_1 to t_N, we learn the parameters of the HMMs using the Baum-Welch algorithm. We then use the emission probability $P(O_{i,N+1}|S_{i,N+1})$ given by the observation-state transition matrix to determine the most likely sequence at t_{N+1} using the Viterbi algorithm. We finally substitute the emission probability $P(O_{i,N+1}|S_{i,N+1})$ from above and $P(S_{i,N+1}|S_{i,N},\mathbf{F}_{i,N})$ from eqn. 23 into eqn. 21 to compute the expectation $E(O_{i,N+1}|O_{i,N},\mathbf{F}_{i,N})$ and get the estimated observed action of u_i: $\hat{O}_{i,N+1}$ (eqn. 20). The details of this estimation can be found in [33].

We now use the estimated social actions $\hat{O}_{i,N+1}(\theta)$ of all users at time slice t_{N+1} to get a set of users who are likely to involve in the diffusion process at t_{N+1} for both the baseline and the attribute social graphs. Next we use G and $\{G(a_k)\}$ to associate edges between the predicted user set, and the users in each diffusion series corresponding to the diffusion collections at t_N. This gives the diffusion collection t_{N+1}, i.e. $\hat{\mathscr{S}}_{N+1}(\theta)$ and $\{\hat{\mathscr{S}}_{N+1;a_k}(\theta)\}$ (ref. section 4.1.4).

4.7 Distortion Measurement

We now compute the diffusion feature vectors $\hat{\mathbf{d}}_{N+1}(\theta)$ or $\{\hat{\mathbf{d}}_{N+1;a_k}(\theta)\}$ based on the predicted diffusion collections $\hat{\mathscr{S}}_{N+1}(\theta)$ and $\{\hat{\mathscr{S}}_{N+1;a_k}(\theta)\}$ from section 4.4. To quantify the impact of attribute homophily on diffusion at t_{N+1} corresponding to $a_k \in \mathscr{A}$, we define two kinds of distortion measures—(1) saturation measurement, and (2) utility measurement metrics.

Saturation Measurement. We compare distortion between the predicted and actual diffusion characteristics at t_{N+1}. The saturation measurement metric is thus given as $1 - D(\hat{\mathbf{d}}_{N+1}(\theta), \mathbf{d}_{N+1}(\theta))$ and $1 - D(\hat{\mathbf{d}}_{N+1;a_k}(\theta), \mathbf{d}_{N+1}(\theta))$, avergaed over all values of $\forall a_k \in \mathscr{A}$ respectively for the baseline and the attribute social graphs. $\mathbf{d}_{N+1}(\theta)$ gives the actual diffusion characteristics at t_{N+1} and $D(A,B)$ Kolmogorov-Smirnov (KS) statistic, defined as $max(|A-B|)$.

Utility Measurement. We describe two utility measurement metrics for quantifying the relationship between the predicted diffusion characteristics $\hat{\mathbf{d}}_{N+1}(\theta)$ or $\{\hat{\mathbf{d}}_{N+1;a_k}(\theta)\}$ on topic θ, and the trends of same topic θ obtained from external time series. We collect two kinds of external trends: (1) *search trends*–the search volume of θ over t_1 to t_{N+1}[8]; (2) *news trends*—the frequency of archived news articles about θ over same period[9]. The utility measurement metrics are defined as follows:

Search trend measurement: We first compute the cumulative distribution function (CDF) of diffusion volume as $E^D_{N+1}(\theta) = \sum_{m \le (N+1)} |l_m(\hat{\mathscr{S}}_{N+1}(\theta))|/Q_D$, where $|l_m(\hat{\mathscr{S}}_{N+1}(\theta))|$ is the number of nodes at slot l_m in the collection $\hat{\mathscr{S}}_{N+1}(\theta)$. Q_D is the normalized term and is defined as $\sum_m |l_m(\hat{\mathscr{S}}_{N+1}(\theta))|$. Next, we compute the CDF of search volume as $E^S_{N+1}(\theta) = \sum_{m \le (N+1)} f^S_m(\theta)/Q_S$, where $f^S_m(\theta)$ is the search volume at t_m, and Q_S is the normalization term. The search trend measurement is defined as $1 - D(E^D_{N+1}(\theta), E^S_{N+1}(\theta))$, where $D(A,B)$ is the KS statistic.

News trend measurement: Similarly, we compute the CDF of news volume as $E^{\mathscr{N}}_{N+1}(\theta) = \sum_{m \le (N+1)} f^{\mathscr{N}}_m(\theta)/Q_{\mathscr{N}}$, where $f^{\mathscr{N}}_m(\theta)$ is the number of archived news articles available from Google News for t_m, and $Q_{\mathscr{N}}$ is the normalization term. The news trend measurement is similarly defined as $1 - D(E^D_{N+1}(\theta), E^{\mathscr{N}}_{N+1}(\theta))$.

[8] http://www.google.com/intl/en/trends/about.html

[9] http://news.google.com/

Using the same method as above, we compute the search and news trend measurement metrics for the attribute social graphs—given as, $1-D(E^D_{N+1;a_k}(\theta), E^S_{N+1}(\theta))$ and $1-D(E^D_{N+1;a_k}(\theta), E^{\mathscr{N}}_{N+1}(\theta))$, averaged over all values of $\forall a_k \in \mathscr{A}$ respectively.

4.8 Experimental Studies

We present our experimental results in this section that validate the proposed framework of modeling diffusion. We utilize a dataset that is a snowball crawl from Twitter, comprising about 465K users, with 837K edges and 25.3M tweets over a time period between Oct'06 and Nov'09. For our experiments, we focus on a set of 125 randomly chosen "trending topics" that are featured on Twitter over a three month period between Sep to Nov 2009. For the ease of analysis, we organize the different trending topics into generalized themes based on the popular open source natural language processing toolkit called "OpenCalais" (http://www.opencalais.com/).

We discuss attribute homophily subject to variations across the different themes, and averaged over time (Oct-Nov 2009). Figure 9 shows that there is considerable variation in performance (in terms of saturation and utility measures) over the eight themes.

In the case of saturation measurement, we observe that the location attribute (LOC) yields high saturation measures over themes related to events that are often "local" in nature: e.g. (1) 'Sports' comprising topics such as 'NBA', 'New York Yankees', 'Chargers', 'Sehwag' and so on–each of them being of interest to users respectively from the US, NYC, San Diego and India; and (2) 'Politics' (that includes topics like 'Obama', 'Tehran' and 'Afghanistan')—all of which were associated with important, essentially local happenings during the period of our analysis. Whereas for themes that are of global importance, such as 'Social Issues', including topics like '#BeatCancer', 'Swine Flu', '#Stoptheviolence' and 'Unemployment', the results indicate that the attribute, information roles (IRO) yields the best performance—since it is able to capture user interests via their information generation and consumption patterns.

From the results on utility measurement, we observe that for themes associated with current external events (e.g. 'Business-Finance', 'Politics', 'Entertainment-Culture' and 'Sports'), the attribute, activity behavior (ACT) yields high utility measures. This is because information diffusing in the network on current happenings, are often dependent upon the temporal pattern of activity of the users, i.e. their time of tweeting. For 'Human-Interest', 'Social Issues' and 'Hospitality-Recreation', we observe that the content creation attribute (CCR) yields the best performance in prediction, because it reveals the habitual properties of users in dissipating information on current happenings that they are interested in.

From these studies, we interestingly observe that attribute homophily *indeed* impacts the diffusion process; however the particular attribute that can best explain the actual diffusion characteristics often depends upon: (1) the metric used to quantify diffusion, and the (2) topic under consideration.

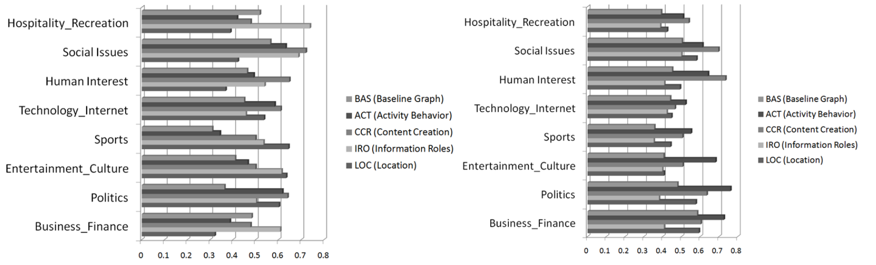


Fig. 9 Mean saturation and utility measurement of predicted diffusion characteristics shown across different themes.

5 Summary and Future Work

Our central research goal in this chapter has been to instrument the three organizing principles that characterize our communication processes online: the information or *concept* that is the content of communication, and the *channel* i.e. the media via which communication takes place. We have presented characterization techniques, develop computational models and finally discuss large-scale quantitative observational studies for both of these organizing ideas

Based on all the outcomes of the two research perspectives that we discussed here, we believe that this research can make significant contribution into a better understanding of how we communicate online and how it is redefining our collective sociological behavior. Beyond exploring new sociological questions, the collective modeling of automatically measurable interactional data will also enable new applications that can take advantage of knowledge of a person's social context or provide feedback about her social behavior. Communication modeling may also improve the automated prediction and recognition of human behavior in diverse social, economic and organizational settings. For collective behavior modeling, the social network can define dependencies between people's behavior with respect to their communication patterns, and features of the social network may be used to improve prediction and recognition. Additionally, some of the statistical techniques developed in this thesis for analyzing interpersonal communication may find new application to behavior modeling (collective or otherwise) and machine learning.

In the future, we are interested in two different non-trivial problems that can provide us with a deeper and more comprehensive understanding of the online communication process. The first of the two problems deals with the idea of evolution of network structure from an ego-centric perspective, in the context of online social spaces that feature multiplex ties. The second problem is geared towards exploring how sociological principles such as homophily (or heterophily) impacts media creation (e.g. uploading a photo on Flickr, or favorting a video on YouTube) on the part of the users. We are interested to study how the observed social interactions among the individuals impact such dynamics. Note, both of the proposed problems consider an observed sociological phenomena prevalent on the social media sites, and attempts to understand it with the help of large-scale quantitative observational studies.

References

1. Lada Adamic and Eytan Adar. How to search a social network. *Social Networks*, 27(3):187–203, July 2005.
2. Eytan Adar and Lada A. Adamic. Tracking information epidemics in blogspace. In *WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 207–214, Washington, DC, USA, 2005. IEEE Computer Society.
3. Eytan Adar, Daniel S. Weld, Brian N. Bershad, and Steven S. Gribble. Why we search: visualizing and predicting user behavior. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 161–170, New York, NY, USA, 2007. ACM.
4. Eytan Bakshy, Brian Karrer, and Lada A. Adamic. Social influence and the diffusion of user-created content. In *EC '09: Proceedings of the tenth ACM conference on Electronic commerce*, pages 325–334, New York, NY, USA, 2009. ACM.
5. Frank M. Bass. A new product growth model for consumer durables. *Management Science*, 15:215–227, 1969.
6. Charles R. Berger and Richard J. Calabrese. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research*, 1(2):99–112, 1975.
7. Ronald S. Burt. Toward a structural theory of action: Network models of social structure, perception and action. *The American Journal of Sociology*, 90(6):1336–1338, 1982.
8. Ronald S. Burt. Structural holes and good ideas. *The American Journal of Sociology*, 110(2):349–399, 2004.
9. Robert B. Cialdini and Noah J. Goldstein. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621, February 2004.
10. James Coleman. *Foundations of Social Theory*. Belknap Press of Harvard University Press, August 1998.
11. Munmun De Choudhury, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. Can blog communication dynamics be correlated with stock market activity? In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 55–60, New York, NY, USA, 2008. ACM.
12. Munmun De Choudhury, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. Social synchrony: Predicting mimicry of user actions in online social media. In *CSE '09: Proceedings of the 2009 International Conference on Computational Science and Engineering*, pages 151–158, Washington, DC, USA, 2009. IEEE Computer Society.
13. Munmun De Choudhury, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. What makes conversations interesting?: themes, participants and consequences of conversations in online social media. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 331–340, New York, NY, USA, 2009. ACM.
14. Scott L. Feld. The focused organization of social ties. *American Journal of Sociology*, 86(5):1015–1035, 1981.
15. Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. Statistical analysis of the social network and discussion threads in slashdot. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 645–654, New York, NY, USA, 2008. ACM.
16. M. S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.
17. Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87, New York, NY, USA, 2005. ACM.

18. Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA, 1999. ACM.
19. David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.
20. Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM.
21. Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, New York, NY, USA, 2008. ACM.
22. Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 915–924, New York, NY, USA, 2008. ACM.
23. D. Liben-Nowell and Jon Kleiberg. Tracing information flow on a global scale using internet chain-letter data. *PNAS*, 105(12):4633–4638, 2008.
24. Miller Mcpherson, Lynn S. Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.
25. Miller McPherson and Lynn Smith-Lovin. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American sociological review*, 52(3):370–379, 1987.
26. Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 101–110, New York, NY, USA, 2008. ACM.
27. Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 533–542, New York, NY, USA, 2006. ACM.
28. Gilad Mishne and Natalie Glance. Leave a reply: An analysis of weblog comments. In *In Third annual workshop on the Weblogging ecosystem*, 2006.
29. Chih-Hui Lai Mor Naaman, Jeffrey Boase. Is it really about me? message content in social awareness streams. In *CSCW '10: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, New York, NY, USA, 2010. ACM.
30. Kevin Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, July 2002.
31. Theodore Mead Newcomb. *The acquaintance process*. Holt, Rinehart and Winston, New York, NY, 1961.
32. Martin Potthast. Measuring the descriptiveness of web comments. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 724–725, New York, NY, USA, 2009. ACM.
33. Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. pages 267–296, 1990.
34. Anne Schuth, Maarten Marx, and Maarten de Rijke. Extracting the discussion structure in comments on news-articles. In *WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management*, pages 97–104, New York, NY, USA, 2007. ACM.
35. Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the evolution of user interaction in facebook. In *WOSN '09: Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42, New York, NY, USA, 2009. ACM.
36. D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296(5571):1302 – 1305, May 2002.
37. Fang Wu and Bernardo A. Huberman. Popularity, novelty and attention. In *EC '08: Proceedings of the 9th ACM conference on Electronic commerce*, pages 240–245, New York, NY, USA, 2008. ACM.