Why Do We Converse on Social Media? An Analysis of Intrinsic and Extrinsic Network Factors

Munmun De Choudhury
School of Communication and Information
Rutgers, The State University of New Jersey
m.dechoudhury@rutgers.edu

Hari Sundaram
School of Arts, Media and Engineering
Arizona State University, Tempe, AZ 85281, USA
hari.sundaram@asu.edu

ABSTRACT

We are motivated in our work by the following question: *what factors influence individual participation in social media conversations?* This question is important due to several reasons. First, conversations around user posted content, is central to the user experience in social media sites, including Facebook, YouTube and Flickr. Second, understanding *why* people participate, can have significant bearing on the following fundamental research questions: social network evolution including changes to the network structure, and information flow.

Our approach is as follows. We first identify several key aspects of social media conversations, distinct from both online forum discussions and other social networks. These aspects include intrinsic and extrinsic network factors. There are three factors *intrinsic* to the network : social awareness, community characteristics and creator reputation. The factors *extrinsic* to the network include: media context and conversational interestingness. We develop one hypothesis for each factor to test the influence of the factor on individual participation. There are two technical contributions of this paper, both related to testing of hypothesis: a Support Vector Regression based prediction framework to evaluate each hypothesis, and a Bayesian Information Criterion (BIC) metric to identify the optimal factor combination. We have interesting findings. First, we show that the factors that influence participation depend on the *media type*: YouTube participation is different from a weblog such as Engadget. Second, different sets of factors influence newcomer and existing participants. Finally, we show that an optimal factor combination improves prediction accuracy of observed participation, by $\sim$9–13% and $\sim$8–11% over using just the best hypothesis and all hypotheses respectively. This reveals that there is likely to be a complex set of factors responsible for the nature of participation observed on different social media conversations today.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems; J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms

Algorithms, Experimentation.

Keywords

Blogs, Conversations, Engadget, Flickr, Huffington Post, Participation, Rich media, Social media, YouTube.

1. INTRODUCTION

Today, rich media sites including Flickr and YouTube as well as weblogs including Engadget and Huffington Post have emerged as popular channels for the expression of individual interests, ideas and opinions[1]. These rich media sites allow users to share content, including uploading images, text and videos. Importantly, the shared content allows users to communicate with other users, through comments on the shared media object. We define a sequence of temporally-ordered comments on the shared media object, as a "conversation."

Conversations are important to understand the nature of the underlying social network [1]. In particular, conversations can be used to study the following: user behavior [2] and information roles, including content dissipators [3], impact on information cascades [4], and influence propagation [5]. Hence, it is important to understand user participation in the context of social media conversations. For example, why do certain conversations exhibit continued and increasing participation from individuals? In this light, our work in this paper is motivated by the following question: *what are the factors that influence individual participation in social media conversations?* Notice that by "participation," we mean that a user has posted comments on a conversation.

Understanding the motivations behind participation of individuals in social media conversations involves several challenges. These challenges are related to key aspects of the social network: the inherent culture of interaction within the greater community, the affinity of the community to invite new individuals, the standard practices of social actions and the goal and purpose of the community-wide interactions. Contemporary online communities support different types of social interaction, and cater to different kinds of audiences. Rich media sites, for example, including YouTube and Flickr, primarily cater to sharing of media objects. On the other hand, blog forums such as Engadget or Huffington Post are directed towards technology-savvy or liberal political audiences who intend to remain engaged in interactions around news events. Therefore, it is likely that different social media sites will have different factors driving conversational participation within their sites. Furthermore, it is likely that there are differences between the motivations of newcomers to participate, compared to the existing members.

[1]As of May 2010, YouTube features more than 2 billion views a day and 24 hours of video uploaded per minute: http://www.digitalbuzzblog.com/infographic-youtube-statistics-facts-figures/

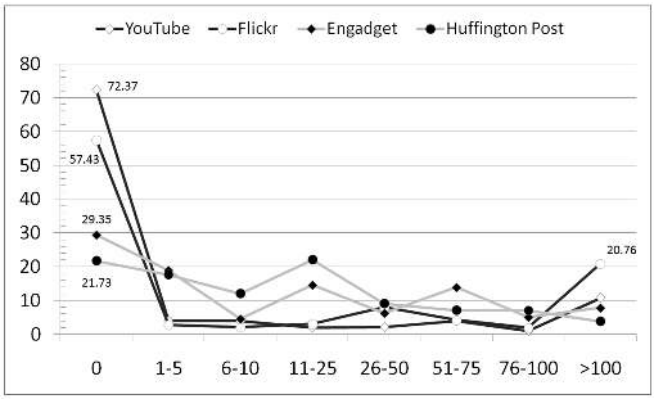

Figure 1: Participation (percentage of users on the y-axis and number of posted comments per user, on the x-axis) on conversations from two types of social media sites: rich media and blog forums. We notice a marked difference in newcomer participation — rich media datasets (YouTube:72.37%, Flickr: 57.43%) attract more newcomers than blog forums (Engadget: 29.35%, Huffington Post: 21.35%). Notice that Flickr has a significant core following (20.76%)—sustained participation from individuals who have posted more than 100 comments.

To establish these differences empirically, we show the distribution of participation (percent) over two types of social media sites: two rich media sites and two blog forum sites in (Figure 1). We notice a marked difference in the nature of user participation — in contrast to blogs, rich media features a large percentage of newcomers. Hence identifying factors influencing participation in each of these sites, and how they vary across the types of sites and participants, is critical. In particular, a careful analysis of participation can help contextualize network phenomena (e.g. distribution of information roles, or network dynamics including changes to the structure and information flow) within these sites. An application of our work includes better design of social media websites — in particular, sites where individuals interact with a shared media object (videos, photos, blogs).

1.1 Our Approach

We define the participation of individuals on a social media conversation as "collective participation." There are two aspects to it: newcomers and existing participants. The former, includes individuals who have not posted a comment or reply on the particular conversation thus far. The latter includes participants who have posted at least one comment or reply at an earlier point in time.

We identify *intrinsic* as well as *extrinsic* network factors influencing collective participation from newcomers and existing participants. The nature of the social network in which the conversation is embedded influences *intrinsic* network factors. Intrinsic network factors include: an individual's 'social awareness,' including peer feedback, 'community characteristics' including the ability to sustain users, and 'reputation' of the media creator. Participants also receive external 'information signals' through *extrinsic* network factors, that may be due to an image/video posted in response to an external event, or associated with emergent themes due to conflicting opinions. The extrinsic factors therefore include the 'media context,' including visual/textual content, tags, and 'conversational interestingness.'

Following these two categories of factors, we develop one hypothesis for each factor to test the influence of the factor on participation. In order to examine how well each hypothesis can be attributed to the observed participation of individuals, we adopt a prediction approach. Our goal is to utilize each hypothesis as a feature in a prediction framework — we perform regression to determine a predicted measure of participation. The better the predicted measure, the more likely the hypothesis is influencing participation.

There can be several ways to qualitatively validate the proposed hypotheses including via ethnographic studies. In this work, however, we adopt a quantitative prediction approach. Prediction of observed participation based on the different hypotheses helps us understand the motivation for an individual to come back to the different types of social media sites. There are two technical contributions of this paper, both related to testing of hypothesis via prediction of observed participation:

- We use a Support Vector Regression based prediction framework to evaluate each hypothesis. Specifically, we test the ability of a factor, including intrinsic and extrinsic factors, to explain observed participation, including newcomers and existing participants.
- We propose a Bayesian Information Criterion (BIC) metric to identify the optimal factor combination. A combination of factors may better explain collective participation.

1.2 Main Results

We tested our hypotheses on two dataset classes — two rich media datasets, Flickr and YouTube, and two blog forum datasets, Engadget and Huffington Post. Our results indicate that different factors influence conversations from the two data classes differently. On one hand, extrinsic network factors, including media context and conversational interestingness, explain participation on rich media conversations. On the other hand, intrinsic network factors, including social awareness and community characteristics seem to explain participation on blog forums. We show that shared media contextual attributes, including visual and textual content, tags etc., influence newcomer participation. In contrast, existing participants seem to rely more on the characteristics of the community for continued participation.

Testing of hypothesis combination also yields insights. Interestingly, we find that including *all* of the intrinsic and extrinsic network factors does not yield the best prediction accuracy. Instead, we note that the optimal combination of factors improves prediction accuracy significantly by $\sim$9–13% and $\sim$8–11% respectively over using just the best factor and all factors. Specifically, we observe that: (a) for rich media conversations, a combination of the extrinsic network factors quantify participation better; and (b) in the case of blog forums, a combination of the intrinsic factors perform the best. This reveals that there is likely to be a complex set of factors responsible for the nature of participation observed on different social media conversations today.

The rest of this paper is organized as follows. In section 2 we discuss prior work. Sections 3, 4 and 5 present the datasets, the different factors behind participation and the framework used to predict it. In section 6 we present our experiments involving the validation of the different hypotheses or factors relating to participation. We test the impact of combining multiple hypotheses in section 7. We conclude with a discussion of some open issues and our contributions in sections 8 and 9 respectively.

2. RELATED WORK

Over several years, sociologists have been interested in understanding individual participation that underpins social movements. Dixon et al. [6] considered aggregate network processes that may

condition the costs and benefits of participation in social movements. Recent work on understanding participation over the Internet has focused on factors associated with continued contribution of individuals on newsgroups, discussion forums, and online communities and networks [7–10]. In the following two paragraphs, we organize the related work towards understanding participation in communities and rich media repositories. This aligns with the organization of the analysis presented in this paper.

Participation in Communities. Lampe et al. [11] examined the participation of users on the technical community Slashdot and substantiated three explanations for participation. They were: learning transfer from previous experiences, observation of, as well as feedback from other participants. Joyce and Kraut [12] studied the factors behind participation in newsgroups. Specifically they considered whether the response received on the first post of an newcomer motivated them to further participate in the communities and observed that the emotional tone of the feedback received had little affect on the individual's motivation to post again. In the context of social networks and social media, Burke et al. [13] studied content contribution on Facebook. They predicted how mechanisms such as social learning, singling out, feedback, and distribution were able to quantify long-term sharing of content (e.g. photos) based on their experiences in the first two weeks.

Participation in Rich Media. In the context of the rich media site Flickr, Nov et al. [10] studied how the tenure in a community affects a variety of participation types for individuals. Negoescu et al [14] adopted a human-centered approach to study two photo-sharing communities: Flickr and Kodak Gallery. Finally, Miller et al [15] studied photography practices, privacy perspectives and socialization styles on Flickr.

Limitations of Prior Work: The state-of-the-art has made significant contributions to understanding factors behind voluntary participation in physical and online communities (e.g. open source forums, Wikipedia, Facebook). A key property of these online communities is the following: there are clear incentives behind an individual's participation in the discussion forum, in editing a Wikipedia article or posting/tagging a photo on Facebook. The incentives could range from contribution to an open source project (in open source forums), generating knowledge (in Wikipedia) or merely the desire to remain posted with one's real world social ties (in Facebook). Therefore, from the prior literature we gain the insight that participation can in these contexts be explained by considering intrinsic factors within the social network. Such factors include, the awareness of a participant to feedback/responses from her peers or her familiarity with the peers in the past.

However, prior research has not investigated participation in the context of the conversations in rich media around which a social network evolves. It is natural to conjecture that *a combination of factors*, such as awareness of the individual to feedback from peers, community behavior as well as conversational interestingness is likely to impact participation. Understanding such factors that influence participation has not received sufficient attention in prior work. Addressing these concerns is a major focus in this work.

3. SOCIAL MEDIA CONVERSATIONS

In this section, we first describe several key social media conversation features, and then provide an overview of the datasets used in this paper.

3.1 Conversations

Social media conversations possess unique characteristics. These features of social media conversations are different from online forum discussions, where user participation has been typically investigated. The key features of conversations include: community, presence of shared media and conversational interestingness.

Community. Shared media conversations can promote cohesive interaction amongst community members. Members of the community can join a specific conversation due to several reasons. First, individuals can come together because they share a common interest in the topic. Second, individuals may be interested in expressing their opinion on a media object related to a recent event. Finally, they may be interested in exchanging ideas with familiar community members, whom they observe participating in the conversation. Conversations also provide a unique framework for individuals to observe peer activity around a specific topic of interest. Such observations can thus be used by a participant to infer characteristics of the larger community around the conversation, that is interested in discussing the particular topic [1]. These community characteristics may include the following: community size, community cohesiveness, including dense cliques, whether there is sustained participation, and if the community can attract and retain new participants. Thus, an individual's observations of the larger community is likely to influence her participation in a conversation.

Shared Media. Social media conversations take place in the context of a shared media object, including a video on YouTube, or a post on the technology blog, Engadget. Naturally, the content of the media object—e.g., visual features of an image/video, textual content of a blog post is likely to impact an individual's desire to participate in the associated conversation. Additionally, each media object typically engenders several contextual attributes, which can attract an individual's attention as well [16], prompting her to post a comment on the conversation. These contextual attributes includes tags, ratings, recency of the media object etc. Hence, analysis of factors behind voluntary participation in these conversations needs to consider the shared media context, including its content.

Conversational Interestingness. Temporal theme evolution is a key characteristic of social media conversations. New themes slowly emerge due to new user comments, and over time, the conversation topic can bear little resemblance to the original conversation topic [17]. In this way, certain themes can emerge to be highly popular. The theme popularity affects the participants who comment in such themes: the participants become important in the context of the conversation. In [17], the authors operationalize temporal evolution of a conversation by the "interestingness" measure of the conversation. We conjecture that the degree of interestingness of a conversation, influences individual participation.

We also consider additional factors: an individual's awareness of peer activities and the reputation and tenure of the media creator in the social network. Such factors are motivated by observations from prior research regarding online participation—self-development, enjoyment [10], reputation [9], feedback and attention [8, 11, 13].

3.2 Datasets

A key goal in this work is to understand the factors affecting collective participation in different types of social media conversations. Social media conversations take place under a variety of contextual conditions. We identify two different conversational contexts: conversations centered around a shared rich media object (image, video) and conversations centered around shared textual content, including blogs. We utilize two datasets from each of the two categories—two rich media websites, Flickr (`http://www.flickr.com/`) and YouTube (`http://www.youtube.com/`), and

Table 1: Description of conversations on different rich media and blog datasets.

DATASET	MEDIA	CONVERSATION
Rich Media Datasets		
YouTube	Video	Comments on the video
Flickr	Photo	Comments on the photo
Blog Forum Datasets		
Engadget	Blog post	Comments on the blog post
Huffington Post	Blog post	Comments on the blog post

two blog forums, Engadget (`http://www.engadget.com/`)[2] and Huffington Post (`http://www.huffingtonpost.com/`)[3].

In Table 1, we provide a summary definition of a conversation in each of the above datasets, along with the shared media object around which we intend to quantify the collective participation.

We describe the details of each dataset in Table 2. All of the datasets were were crawled for research purposes using their respective APIs. Additionally, the time spans of these crawled datasets have been chosen to be of approximately the same length ($\sim$ 147 days) and are given as follows. YouTube: Sep 1, 2008–Jan 31, 2009; Flickr: Feb 1–Jun 30, 2008; Engadget: Apr 1, 2008–Aug 31, 2008; and Huffington Post: May 15, 2008–Oct 10, 2008.

Table 2: Details of the four datasets.

Dataset	#Participants	#Conversations	#Comments
Rich Media Datasets			
YouTube	17,736,361	272,810	145,682,273
Flickr	4,304,525	305,258	26,557,446
Blog Forum Datasets			
Engadget	78,740	45,073	6,580,256
Huff Post	59,282	24,479	4,748,837

4. FACTORS IN SOCIAL PARTICIPATION

There are several factors that can affect the degree of participation in social media conversations — both for the newcomers and existing participants. We categorize them as intrinsic and extrinsic network factors. Factors *intrinsic to the network* include social awareness, community characteristics and creator reputation. Factors *extrinsic to the network* are features related to media context and conversational interestingness. In the rest of this section, we propose several features for intrinsic and extrinsic factors.

4.1 Intrinsic Network Factors

Social Awareness. Participation of individuals in social media conversations is dependent upon factors that induce social awareness in an individual. The factors that influence social awareness in a conversational setting in digital communities have been studied in prior literature [8, 9, 11–13]. We utilize three measures of social awareness:

- *Familiarity:* We quantify the degree of *familiarity* with participants with respect to an individual associated with a conversation to be a function of the number of times they co-participated in any prior conversation on the same topical category. For the n-th conversation, familiarity $F_K^{(n)}$ at time slice t_K is thus given by the ratio of the mean frequency of co-participation of every participant u with every other participant v on prior conversations, to the mean frequency of participation of all participants in all prior conversations between t_1 and t_K.

- *Feedback:* Next we quantify the degree of *feedback* with participants with respect to an individual associated with a conversation to be a function of the number of replies she receives from other participants. For the n-th conversation, feedback $D_K^{(n)}$ at time slice t_K is thus given by the mean number of replies each participant in the conversation receives, to the number of comments / replies s/he has had posted until t_K.

- *Dialogue:* Presence of *dialogue* $L_K^{(n)}$ among the participants in the n-th conversation is given by the ratio of the frequency of all the replies to frequency of all the comments until t_K. It is therefore a measure of the overall back and forth communication (comment/reply) has happened between the participants in the past.

HYPOTHESIS 1. *Collective participation on a social media conversation is affected by the degree of social awareness of the participating individuals, including their familiarity with other participants in the conversation, feedback from others and dialogue among others.*

Community Characteristics. Properties of the overall community also influence collective participation in conversations. We consider a community to be a set of individuals who engage in commentary centered around a broad topic. A typical community in our dataset, for example, on YouTube: a set of individuals who write comments or replies around shared videos on the topic of "News & Politics".

We consider different properties, structural and temporal, to characterize online communities: community size, community activity, community cohesiveness and community sustenance. We describe each of these characteristics below:

- *Community size* S_K at a certain time slice t_K is defined as the number of unique individuals who have posted a comment or a reply at least once on all conversations associated with media objects belonging to a certain topical category.

- *Community activity* A_K at a certain time slice t_K is the mean degree of activity of the individuals in a community. It is given by the mean number of postings of comments and replies across all the individuals in the community.

- *Community cohesiveness* H_K at a certain time slice t_K is defined as the mean clustering coefficient of the communication graph. The graph is induced by the co-participation of individuals commenting or replying to all conversations associated with media objects belonging to a certain topic. Notice that the communication graph is an undirected weighted

[2]Engadget is a technology weblog and podcast about consumer electronics. The blog is usually updated multiple times a day with articles on gadgets and consumer electronics, typically by an Editor. It also posts rumors about the technological world, frequently offers opinion within its stories, and produces profuse commentary from registered users centered around the stories.

[3]Huffington Post is an American news website and aggregated blog featuring various news sources and columnists. The site offers coverage of politics, media, business, entertainment, living, style, the green movement, world news, and comedy, and is a top destination for news, blogs, and original content. Huffington Post has an active community, with over one million comments made around the posted blog stories.

Table 3: Media context on multiple rich media and blog datasets.

DATASET	MEDIA CONTENT FEATURES	MEDIA META-DATA
	Rich Media Datasets	
YouTube	Visual features of the video—color (color histogram, color moments), texture (GLCM, phase symmetry) [18, 19], shape (radial symmetry, phase congruency) [20, 21] and keypoint location features (SIFT) [22]. These features are computed over a key frame in each video, where the key frame corresponds to the one at the median time of the duration of the video	Number of views[a], number of 'favorites'[a], ratings, number of linked sites, time elapsed since video upload (recency), video duration
Flickr	Visual features of the photo—color (color histogram, color moments), texture (GLCM, phase symmetry) [18, 19], shape (radial symmetry, phase congruency) [20, 21] and keypoint location features (SIFT) [22].	Number of tags[a], number of notes, number of views[a], number of 'favorites'[a], number of associated groups, time elapsed since photo upload (recency)
	Blog Forum Datasets	
Engadget	tf-idf (term frequency-inverse document frequency) based features of the blog content; where the content is represented as a stemmed and stop-word eliminated bag-of-words	Number of tags[a], time elapsed since blog was posted (recency), number of Facebook "likes"[a], length of the post
Huffington Post	tf-idf (term frequency-inverse document frequency) based features of the blog content; where the content is represented as a stemmed and stop-word eliminated bag-of-words	Number of tags[a], time elapsed since blog was posted (recency), number of Facebook "likes"[a], length of the post

[a] Variable is log-transformed to correct for skew.

graph $G(V, E; t_K)$ where the nodes V are the individuals who have posted a comment or a reply at least once on all conversations associated with media objects belonging to a certain topical category until t_K. An edge $e \in E$ exists between two individuals in V if they have commented/replied together (i.e. co-participated) on the same conversation belonging to the topical category at least once until t_K. The weight on the edge is proportional to the mean frequency the co-participation between t_1 and t_K.

- *Community sustenance* U_K at a certain time slice t_K is defined as the mean degree of retention of communicating individuals over time. Sustenance is a function of the number of individuals who repeatedly return to the community over time to post comments / replies on conversations belonging to the particular topic. For a community $\mathcal{C}_K$ at a certain time slice t_K sustenance is defined as follows:

$$U_K = \frac{1}{K-1} \sum_{m=1}^{K-1} \frac{|\mathcal{C}_K \bigcap \mathcal{C}_m|}{|\mathcal{C}_K|}, \tag{1}$$

where $\mathcal{C}_m$ is the community at time slice t_m.

HYPOTHESIS 2. *Collective participation on a social media conversation is affected by the characteristics of the larger community, including its size, how active and cohesive its members are, and to what degree it is able to sustain its members over time.*

Creator[4] Reputation. Since social media conversations are typically centered around a media object, the identity or characteristics of the creator is likely to play an important role in the communication. Let the reputation be defined as the one-dimensional vector $\mathbf{R}_K^{(n)}$ corresponding to the n-th conversation at time slice t_K. The vector contains measures of the following attributes: number of media objects uploaded / posted by the individual until t_K, his or her number of (social) contacts in the community until t_K (if applicable), i.e. his or her authority measure in the network[5], and the duration of his or her 'tenure' i.e. the time elapsed until t_K, since the date s/he joined the website [10].

HYPOTHESIS 3. *Collective participation on a social media conversation is affected by the reputation of the creator of the associated media artifact, including his or her activity in media creation, his network authority score and tenure in the larger community.*

4.2 Extrinsic Network Factors

Media Context. As mentioned earlier, a distinct feature of participation on social media conversations is that it takes place around a shared media object. Hence the media context is also useful in analyzing the degree of collective participation over time. We consider two kinds of collective participation media contexts: the visual/textual content (features) of the media object, and media meta-data. A detailed description of the two different aspects of the media context is described in Table 3. Corresponding to the n-th conversation and at time slice t_K, we therefore assume that the visual / content features are denoted as $\mathbf{V}_K^{(n)}$ and the meta-data features as $\mathbf{M}_K^{(n)}$ — note that both of these aspects are one-dimensional feature vectors.

HYPOTHESIS 4. *Collective participation on a social media conversation is affected by the context associated with the media artifact, including its visual or textual content as well as media meta-data, including its ratings, views, tags and recency of upload.*

Conversational Interestingness. A typical aspect of social media conversations is that they engender communication around the shared media spanning a variety of external events. As a result, *we conjecture that collective participation will be significantly affected by the evolving nature of the conversation itself.* We consider a subjective temporal property of the conversations: known as "interestingness". We utilize the interestingness model proposed in [17] to compute this measure as a real scalar value in the range

[4] A creator is simply the individual who uploads a video on YouTube, shares a photo on Flickr or write blog posts on Engadget or Huffington Post.

[5] Variable is log-transformed to correct for skew.

[0,1]. Interestingness of a conversation at any given time depends on its themes (popular themes featured in a conversation are likely to make it interesting to individuals and facilitate participation); and also the prior communication activity of its participants.

Since conversational interestingness is a new feature, we briefly review its calculation. Specifically, the authors [17] propose the following steps. First, conversational themes are detected using a temporally regularized mixture of multinomials model. Second interestingness of participants and interestingness of conversations are determined based on an one-dimensional random walk model. Finally, a joint optimization framework of interestingness is used to effectively compute interestingness, that incorporates temporal smoothness constraints. In this work, we denote this interestingness measure of the n-th conversation at time slice t_K as $I_K^{(n)}$.

HYPOTHESIS 5. *Collective participation on a social media conversation is affected by the characteristics of the conversation itself, such as its interestingness over time, where interestingness is characterized by the popularity of the conversational themes and the communication properties of the participants around those themes.*

We now provide a brief summary of the various factors behind participation that have been proposed in this section. We proposed a number of intrinsic and extrinsic network factors that are likely to impact participation of newcomers and existing participants in social media conversations. These include three intrinsic factors: social awareness, community characteristics and creator reputation. While the two extrinsic factors are: media context and conversational interestingness.

5. A PREDICTION FRAMEWORK

In this section, we propose a prediction approach to evaluate each hypothesis in explaining observed participation. In particular, we are interested in explaining participation for newcomers and existing users, in different types of social media conversations (see Figure 1). The goal is to identify the sets of factors which are more effective (i.e. higher accuracy in prediction) in accounting for the observed participation in social media conversations.

Our prediction framework uses a learning framework, that regresses over past degrees of participation using the various factors impacting participation (by treating them as features). Then it predicts the measure of participation at a future point in time by using the best fit coefficients. In this work, we utilize an incremental Support Vector Regression model [23] to predict the degree of observed participation, that can be attributed to each of the five different types of factors.

We begin with by constructing our "ground truth" for quantifying the influence of each type of factor towards newcomer and existing user participation. Let $\mathcal{N}_K^{(n)}$ be the number of comments that are generated on the n-th conversation at time slice t_K, by individuals who had not posted any comments (or replies) on the same conversation between t_1 and t_{K-1}. Furthermore, let $\mathcal{E}_K^{(n)}$ be the number of comments that are generated on the same n-th conversation at time slice t_K, by individuals who had posted at least one comment (or reply) on the same conversation between t_1 and t_{K-1}. In the same way, we determine the degrees of participation over all time slices between t_1 and t_K. Without loss of generality, let us denote a participation vector to be $\mathbf{Y}_{1:K}^{(n)} \in \mathbb{R}^{K\times 1}$.

Next we define five different feature sets, corresponding to the five categories of factors discussed in the previous sub-section. Let, for the n-th conversation, $\mathbf{f}_K^{(n)} \in \mathbb{R}^{1\times d}$ denote the feature vector corresponding to any of these five categories at time slice t_K; d being the number of features (or dimensionality) within the chosen category. The feature vectors can be similarly constructed for all time slices from t_1 to t_{K-1}. Let us represent the matrix of the feature vectors for all time slices between t_1 and t_K as $\mathbf{X}_{1:K}^{(n)} \in \mathbb{R}^{K\times d}$.

We use the data over the first p time slices (where $p < K$) to predict the number of comments from newcomer and existing participants. We split the ground truth vector (or the dependent variable) $\mathbf{Y}_{1:K}^{(n)}$, as well as the feature matrix (or the independent variables) $\mathbf{X}_{1:K}^{(n)}$ into training and testing sets. The first p slices form the training set, while the remaining $p+1$ to K time slices are the test set. The training phase of the SV Regression model (based on a Gaussian RBF kernel), gives us the support vectors, and the best-fit regression coefficients. These coefficients are thereafter applied on the test set over time slices $p+1$ to K to get the predicted measures of participation over time slices $p+1$ through K. The effectiveness of the chosen feature set category is therefore given by the mean percentage accuracy in predicting the value $\hat{\mathcal{N}}_i^{(n)}$ and $\hat{\mathcal{E}}_i^{(n)}$ against the actual values $\mathcal{N}_i^{(n)}$ and $\mathcal{E}_i^{(n)}$ for all time slices $p+1 \leq i \leq K$. The accuracy measure is given as the ratio of the absolute difference between predicted and actual values to the actual value of each type of participation (the number of comments from newcomer and existing participants).

6. VALIDATING HYPOTHESES

We conduct elaborate experimental studies on all the four datasets introduced in section 3, in order to find empirical grounding on the five different hypotheses behind collective participation proposed in this paper. For all the four datasets, we choose the first 97 days ($\sim 65\%$) as the training phase and the next 50 days ($\sim 35\%$) as test set in each case. We avoid using larger training set sizes to prevent overfitting in the prediction task.

6.1 Prediction Performance

We begin by presenting prediction performance of using different feature set categories in accordance with the different hypotheses framed in section 4. The performance is evaluated based on the corresponding percent accuracy metric (discussed in section 5) and we present the results for both dataset types, as well as for newcomer participation as well as that from existing participants (Figure 2).

Rich media vs. Blog Forums. We observe differences in the feature sets that yield the best prediction performance across the two dataset types. For rich media data, extrinsic network factors (media context and conversational interestingness; mean accuracy $\sim$80%) seem to better predictors of participation compared to social awareness and community characteristics. This is because the nature of the shared media is central to triggering users to participate in conversations. For blog forums data, intrinsic network factors (social awareness and community characteristics; mean accuracy $\sim$78%) seem to better predictors of participation compared to the others. This is because participation on these websites are often driven by personal opinions on technology or political happenings. Hence the overall community's response and behavior to a certain event are likely to be important factors behind participation.

Newcomers vs. Existing Participants. There are also significant differences across the factors that affect participation in newcomers and existing participants. Conversational interestingness and community characteristics perform relatively better for all datasets in the case of existing participants. This is because over time they are able to 'learn' a community's dynamics: its nature of activity as well as can judge better (via comparison) the interestingness of the

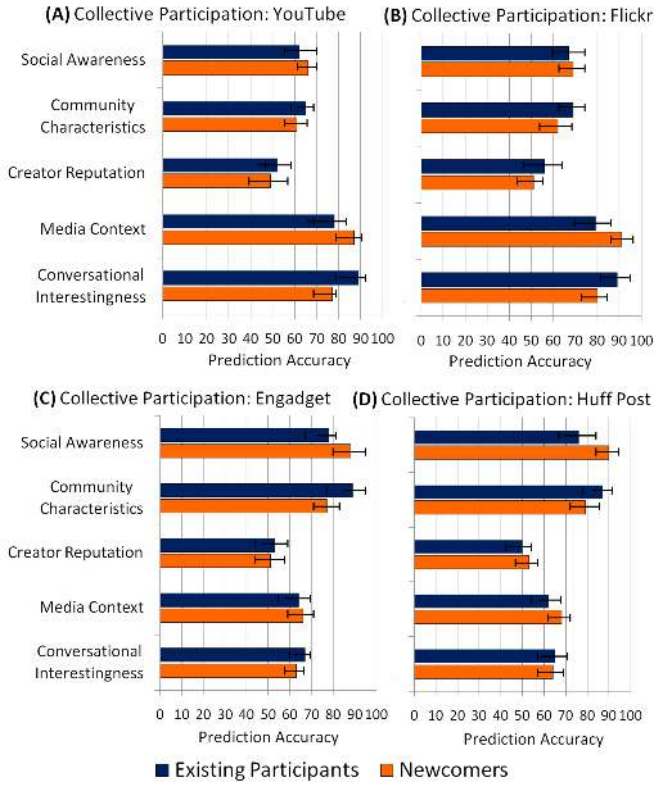

Figure 2: Prediction accuracies (higher numbers are better) of collective participation in social media conversations over four different datasets; corresponding error bars are also shown to illustrate the deviations. Media context and conversational interestingness perform better for rich media datasets, while social awareness and community characteristics perform better for blogs. Creator reputation appears to perform relatively poorly over all datasets.

on-going conversations. Newcomers seem to rely more on media context and social awareness. This is because their participation is likely to be triggered by the popularity of the media shared, or by how the rest of the participants are reacting to their comments.

Creator Reputation. The creator reputation feature does not explain collective participation well for any of the datasets (accuracy ∼49%). We believe that there are two explanations: large number of authors in rich media sites and restrictive media authorship in blogs. Anyone can upload a media object in rich media websites. Since there is no restriction on *who* can upload—the number of creators on rich-media sites is very large. This overabundance of creator choice, in rich media sites, makes the creator a less likely candidate as the *sole* attribute on which to filter media. On the two blog forums analyzed in this work, only a fixed number of Editors can create content. Since all the content on these two blogs are created by editors, the reputation of the editor makes little difference to user participation.

6.2 Statistical Significance

From the results in Figure 2 we observe that that there are differences in prediction performances for the different feature set categories, or hypotheses and the datasets. In order to substantiate the differences, we perform some tests of statistical significance (one-tailed paired Student's t-test) on the prediction accuracy measure of each hypothesis, including both for newcomers and existing participants (ref. Figure 2). In particular, we are interested to investigate if the performances of the "best" hypotheses are statistically significant compared to that of the others.

Table 4: Statistical significance (based on one-tail paired t-test; significance level of 0.05) of the two best performing factors for each dataset, compared to the other features, for same datasets. Here, SA: Social Awareness, CC: Community Characteristics, CR: Creator Reputation, MC: Media Context and CI: Conversational Interestingness. p-values below the significance level are shown in italics. We observe that in each dataset, the performances of the "best" factors are statistically significant compared to the other factors. Note that because we used a one-tail paired t-test, the significance results of X & Y is different from Y & X.

	t	p		t	p
RICH MEDIA DATASETS					
YouTube: df= 272,810; best performing factors: CI, MC					
SA & CI	-926.3	*0.013*	SA & MC	-909.1	*0.028*
CC & CI	-937.9	*0.007*	CC & MC	-924.6	*0.016*
CR & CI	-959.7	*0.001*	CR & MC	-951.7	*0.002*
MC & CI	-892.3	0.116	CI & MC	-884.4	0.204
Flickr: df= 305,258; best performing factors: CI, MC					
SA & CI	-981.5	*0.011*	SA & MC	-914.5	*0.031*
CC & CI	-1035.1	*0.006*	CC & MC	-1052.6	*0.014*
CR & CI	-1263.3	*0.001*	CR & MC	-1352.2	*0.003*
MC & CI	-835.4	0.121	CI & MC	-862.7	0.193
BLOG FORUM DATASETS					
Engadget: df= 45,073; best performing factors: SA, CC					
CC & SA	-318.4	0.147	SA & CC	-298.6	0.192
CR & SA	-451.3	*0.001*	CR & CC	-431.4	*0.002*
MC & SA	-405.7	*0.008*	MC & CC	-424.2	*0.007*
CI & SA	-362.8	*0.009*	CI & CC	-379.1	*0.008*
Huffington Post: df= 24,479; best performing factors: SA, CC					
CC & SA	-224.3	0.091	SA & CC	-183.6	0.075
CR & SA	-373.5	*0.002*	CR & CC	-278.3	*0.001*
MC & SA	-285.8	*0.006*	MC & CC	-237.3	*0.003*
CI & SA	-194.7	*0.003*	CI & CC	-207.5	*0.002*

We consider all the conversations in each dataset. Our experimental setup consists of a one-tail paired t-test that compares the prediction accuracies of the two best performing methods/hypotheses for each dataset (obtained from Figure 2), with that using each of the other methods. Our null hypothesis is that the accuracy measures are sampled from the same distribution and hence the distribution of accuracies over the two methods under consideration would have similar means and therefore account for little differences. We predict that for the two best performing methods in the case of each dataset, the null hypothesis will be false, because the differences in performances of these two methods against others are significantly better.

The measures of the t-statistic and the corresponding p-values for each of the comparisons is given in Table 4. We use the following abbreviations here — SA: Social Awareness, CC: Community Characteristics, CR: Creator Reputation, MC: Media Context and CI: Conversational Interestingness. Notice that the since the test is *one-tail*, the results are not symmetric. Consider, as an example, the attributes CI and MC. In the one-tailed case, comparing MC to

Table 5: Summary of results in prediction of collective participation by newcomers and existing participants. Here SA: Social Awareness, CC: Community Characteristics, CR: Creator Reputation, MC: Media Context and CI: Conversational Interestingness.

	Support–Rich Media	Support–Blog Forums
Newcomers		
SA	Less (-33%)	*High (-11%)*
CC	Less (-37%)	Moderate (-20%)
CR	Minimal (-49%)	Minimal (-47%)
MC	*High (-19%)*	Less (-34%)
CI	Moderate (-28%)	Less (-31%)
Existing participants		
SA	Less (-35%)	Moderate (-24%)
CC	Less (32%)	*High (-15%)*
CR	Minimal (-48%)	Minimal (-51%)
MC	Moderate (-26%)	Less (-36%)
CI	*High (-18%)*	Less (-33%)

CI will be different from the result from comparing CI to MC.

We show the results for the two rich media and two blog forum datasets. In the case of YouTube and Flickr, the results reveal that the best performing methods, i.e. CI and MC yield p-values below the significance level of 0.05, with respect to SA, CC and CR. The same is true for the two best performing methods SA and CC for Engadget and Huffington Post. Consequently, we reject the null hypothesis that there is no difference in the different factors.

6.3 Summary of Findings

We conclude that *not* all the stated hypotheses are able to quantify the observed participation equally well. There are differences across the two classes of dataset, rich media and blog forums, as well as between the participation from newcomers and existing participants. We summarize these findings in Table 5. For the purpose of easy comprehensibility, we indicate how much *support* each hypothesis provides towards quantifying the participation from newcomers and existing participants separately. Support is defined as the negative of the error—the difference between the predicted accuracy and the ground truth observed participation. At zero error, we have the highest support. We define following terms for the support (S): High ($-20 \leq S \leq 0$), Moderate ($-30 \leq S \leq -20$), Less ($-40 \leq S \leq -30$), and Minimal ($S \leq -40$).

7. COMBINING MULTIPLE HYPOTHESES

Collective participation in online media will typically be manifested due to a collection of factors, rather than a single factor, as discussed in the previous section. In this section, we therefore use a Bayesian Information Criterion (BIC) based measure to determine the optimal set of factors to explain collective participation.

7.1 BIC Measure

Our goal is to determine an "optimal" number of factors, typically smaller than the set of all factors, that can best explain the observed participation. This problem can be reduced to a model selection problem, where we fit a model in a learning task, with a number of parameters. In our case, the different factors can be considered as the model parameters, and we are interested in identifying the optimal parameter combination that helps explain the observed participation. We utilize a measure frequently used in model selection—known as Bayesian Information Criterion (BIC) [24] to find the optimal factor subset.

We develop an iterative approach to determine the optimal hypothesis combination using the BIC measure. We start with a random hypothesis, and sequentially add hypotheses to it. The feature vector corresponding to the chosen starting hypothesis is used to predict the collective participation (of newcomers and existing participants) using the Support Vector Regression technique discussed in Section 5. Using the prediction error, we then compute the BIC measure of the combination at the current iteration. Then at the next step we add a hypothesis. The optimal hypothesis added at each iteration, is the one that *minimizes the Bayesian Information Criterion (BIC)* measure [24] in the combination, for predicting participation. This procedure terminates when there are no more hypothesis are left to be added.

This procedure is repeated for different starting seeds. The particular combination that yields the minimum BIC value overall is chosen as the optimal hypothesis combination.

7.2 Results

We present the results of combining hypotheses to predict collective participation in Figure 3. The figure has two parts. In the top part, we show a visual representation of which hypotheses were chosen at each iteration for each starting seed hypothesis. This is shown using linear paths between the hypotheses (each path has a different starting seed hypothesis). That is, in the figure, in the prediction of newcomer participation for YouTube, at iteration 3, for the starting seed hypothesis CI, we have the optimal combination as {CI, MC, SA}, shown with the green dotted path. Next in the bottom part of the figure, we show the BIC value of each hypothesis combination at each iteration (shown in a line plot with the same color and style as the corresponding path in the top part).

The results indicate that combining hypotheses does *indeed* appear to improve the prediction of collective participation for both newcomers and existing participants. It appears that the combinations that perform the best are the ones which have the starting seed as the best performing hypothesis in Figure 2. However, surprisingly enough, using all information in terms of all five hypotheses *does not* yield the best prediction. In fact the best performance, as seen in the BIC curves in Figure 3 are given by hypotheses combinations in the middle of the curve—that is, a selective few hypotheses, on combination, quantify collective participation in the best manner.

Table 6: Summary of results of combining hypotheses in prediction of collective participation.

Dataset	Newcomers	Existing Participants
YouTube	{MC, CI, SA}	{MC, CI, CC}
Flickr	{MC, CI, SA}	{MC, CI, CC}
Engadget	{SA, CC, MC}	{SA, CC, CI}
Huff Post	{SA, CC, MC}	{SA, CC, CI}

A summary of the best performing combinations is shown in Table 6. We also present in Table 7. the prediction accuracies for these best performing combinations and compare them to those of using just the best performing hypothesis (ref. Figure 2) and the combination of all five hypotheses. The best combination improves prediction accuracy significantly by ∼9–13% and ∼8–11% respectively over using just the best hypothesis and all hypotheses.

The combinations that work best (see Table 6), for example, for rich media are ones which utilize extrinsic network factors: MC and CI. For blogs, intrinsic network factors SA and CC play a key

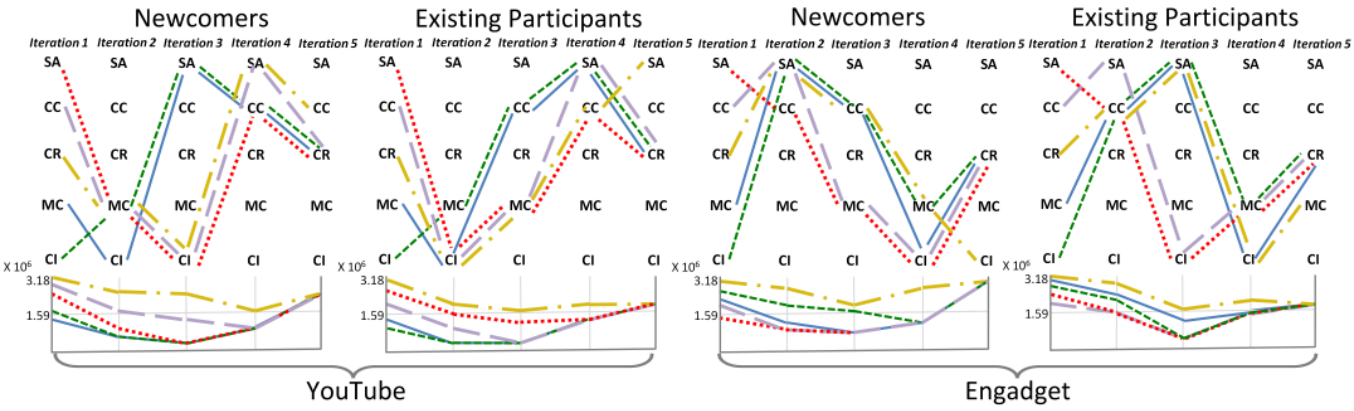

Figure 3: Performance of combining different feature categories (or hypotheses) in predicting collective participation. For each starting feature set, we show which feature sets were selected at each iteration, that minimizes the BIC. The plot at the bottom shows the actual BIC measures of the combinations at each step (lower BIC values are better). Here SA: Social Awareness, CC: Community Characteristics, CR: Creator Reputation, MC: Media Context and CI: Conversational Interestingness.

Table 7: Prediction accuracies using (I) just the best performing hypotheses (Figure 2), (II) optimal hypotheses combination (Figure 3), and (III) all five hypotheses.

Dataset	Newcomers			Existing participants		
	I	**II**	III	I	**II**	III
YouTube	79%	**88%**	80%	80%	**92%**	81%
Flickr	82%	**92%**	82%	80%	**91%**	82%
Engadget	76%	**89%**	78%	76%	**87%**	76%
Huff Post	83%	**93%**	82%	81%	**90%**	80%

role. For newcomers MC and SA are important across both blogs and rich media, while for existing participants, CC and CI are key across all datasets. As before, the creator reputation least affects the participation measures. Hence its inclusion make the prediction worse by increasing the BIC of the combination.

It is reasonable to conclude from these experiments that collective participation on social media conversations are guided by *a complex set of factors*. However, different factors dominate depending on the type of the site: rich media site participation depends more on the properties of the conversation itself, while the blog forums particpation are guided by the social attributes including awareness and community behavior.

8. OPEN ISSUES

We now discuss some of the open issues in this paper. These include the puzzling lack of influence of creator reputation, incompleteness of factors, and the variety of participatory mechanisms.

We observed with some surprise that the factor relating to creator reputation barely influenced collective participation of individuals. There may be several reasons why this was so. First, since these communities are very large, it is likely that there is little awareness of the content creator identity. This is in contrast to smaller well-knit communities, where individuals are aware of media creation activities of fellow community members. Peer awareness, can become one important variable in development of creator reputation. Finally, while we had used a measure of reputation motivated from prior work [10], it may be worthwhile investigating the measure carefully, and developing new measure(s), that can contain additional factors.

It is possible that we have not exhaustively examined the set of factors influencing participation. Unobserved variables, including participant demographics, gender, age, location and cultural norms may also affect participant behavior. Additionally, the participation behavior might evolve over time, in the case of the existing participants. Sentiment and writing style may also influence participation. Fleshing out more extensive factors driving participation remains a ripe area for future research.

Additionally, we have considered only one kind of participation on social media sites: posting comments on conversations. An individual may also participate in other ways: individuals can participate by rating comments, sharing posts and comments of interest. We would be interested to see in future work, if our intrinsic and extrinsic factors can explain these other forms of participation.

Finally, participation on a social media website may also be affected by an individual's intrinsic or idiosyncratic behavior. As an auxiliary method of validation of the factors considered here, ethnographic studies may also be conducted in the future to capture individual behavior that lie beyond the scope of these factors.

9. CONCLUSIONS

In this paper, we investigated several factors to explain participation in social media conversations. Investigating the factors allows us to understand the nature of the underlying social network, including network structure and evolution, and information roles, and influence propagation. Efficient design of social media sites is one potential application of our work.

Our approach was as follows. We first identified several key intrinsic and extrinsic network factors influencing social media conversations, distinct from both online forum discussions and other social networks. We identified three intrinsic factors: social awareness, community characteristics and creator reputation. Furthermore, we identified two extrinsic factors: media context and conversational interestingness. We developed one hypothesis for each factor to test the influence of the factor on individual participation. We developed a Support Vector Regression based prediction framework to evaluate each hypothesis, and a Bayesian Information Criterion (BIC) metric to identify the optimal factor combination.

Our approach addressed two limitations of prior work. First, prior work typically looked at intrinsic network factors affecting the awareness of the participant, including peer familiarity and peer feedback. While these factors are of value and explored in this work, prior work paid little attention to extrinsic network factors, including conversational dynamics and content. We incorporated

both intrinsic and extrinsic factors in our work. A second difference is that that we investigated how a combination of factors influence participation on social media conversations

We presented three interesting findings. First, we showed that extrinsic network factors significantly affected conversations on rich media, while intrinsic network factors were a significant factor for blog forums. Second, awareness of responses and community feedback affected newcomer participation in a conversation. In contrast, the behavior of the overall community, including community support for cohesive participation, or ability of the community to sustain participation, could better explain existing participant behavior. Finally, interestingly enough, we showed that an optimal factor combination improved prediction accuracy of observed participation by $\sim$9–13% and $\sim$8–11% over using just the best hypothesis and all hypotheses respectively.

We plan to investigate several research directions in the future, including a careful analysis of creator reputation, increasing the number of factors that may influence participation, and examining other forms of participation in social media conversations.

10. REFERENCES

[1] Fabricio Benevenuto, Fernando Duarte, Tiago Rodrigues, Virgilio A.F. Almeida, Jussara M. Almeida, and Keith W. Ross. Understanding video interactions in youtube. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, pages 761–764.

[2] Fan Qiu and Yi Cui. An analysis of user behavior in online video streaming. In *Proceedings of the international workshop on Very-large-scale multimedia corpus, mining and retrieval*, VLS-MCMR '10, pages 49–54.

[3] Joan-Isaac Biel and Daniel Gatica-Perez. Wearing a youtube hat: directors, comedians, gurus, and user aggregated behavior. In *Proceedings of the seventeen ACM international conference on Multimedia*, MM '09, pages 833–836.

[4] Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 721–730, New York, NY, USA, 2009. ACM.

[5] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15, New York, NY, USA, 2008. ACM.

[6] Marc Dixon and Vincent J. Roscigno. Status, networks, and social movement participation: The case of striking workers. *The American J. of Sociology*, 108(6):1292–1327, May 2003.

[7] Gerard Beenen, Kimberly Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, and Robert E. Kraut. Using social psychology to motivate contributions to online communities. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, CSCW '04, pages 212–221, New York, NY, USA, 2004. ACM.

[8] Joon Koh, Young-Gul Kim, Brian Butler, and Gee-Woo Bock. Encouraging participation in virtual communities. *Commun. ACM*, 50:68–73, February 2007.

[9] Oded Nov, David Anderson, and Ofer Arazy. Volunteer computing: a model of the factors determining contribution to community-based scientific research. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 741–750, New York, NY, USA, 2010. ACM.

[10] Oded Nov, Mor Naaman, and Chen Ye. Analysis of participation in an online photo-sharing community: A multidimensional perspective. *J. Am. Soc. Inf. Sci. Technol.*, 61:555–566, March 2010.

[11] Cliff Lampe and Erik Johnston. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, GROUP '05, pages 11–20, New York, NY, USA, 2005. ACM.

[12] Elisabeth Joyce and Robert E. Kraut. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11:2006, 2006.

[13] Moira Burke, Cameron Marlow, and Thomas Lento. Feed me: motivating newcomer contribution in social network sites. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 945–954, New York, NY, USA, 2009. ACM.

[14] Radu Andrei Negoescu, Alexander C. Loui, and Daniel Gatica-Perez. Kodak moments and flickr diamonds: how users shape large-scale media. In *Proceedings of the international conference on Multimedia*, MM '10, pages 1027–1030, New York, NY, USA, 2010. ACM.

[15] Andrew D. Miller and W. Keith Edwards. Give and take: a study of consumer photo-sharing culture and practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '07, pages 347–356.

[16] Lyndon Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *Proceedings of the 15th international conference on Multimedia*, MM '07, pages 631–640, New York, NY, USA, 2007. ACM.

[17] 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.

[18] Zhitao Xiao, Zhengxin Hou, Changyun Miao, and Jianming Wang. Using phase information for symmetry detection. *Pattern Recogn. Lett.*, 26:1985–1994, October 2005.

[19] Munmun De Choudhury, Hari Sundaram, Yu-Ru Lin, Ajita John, and Dorée Duncan Seligmann. Connecting content to community in social media via image content, user tags and user communication. In *ICME 2009. IEEE International Conference on Multimedia and Expo*, pages 1238 – 1241.

[20] Gareth Loy and Alexander Zelinsky. Fast radial symmetry for detecting points of interest. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25:959–973, August 2003.

[21] Zheng Liu and Robert Laganière. Phase congruence measurement for image similarity assessment. *Pattern Recogn. Lett.*, 28:166–172, January 2007.

[22] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004.

[23] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, 1998.

[24] Scott S. Chen and P. S. Gopalakrishnan. Clustering via the bayesian information criterion with applications in speech recognition. In *ICASSP '98: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 645–648. IEEE, 1998.