Contextual Prediction of Communication Flow in Social Networks

Munmun De Choudhury Hari Sundaram Ajita John Dorée Duncan Seligmann
Arts Media & Engineering, Arizona State University *Collaborative Applications Research, Avaya Labs*
Email: {munmun.dechoudhury,hari.sundaram}@asu.edu, {ajita,doree}@avaya.com

Abstract

The paper develops a novel computational framework for predicting communication flow in social networks based on several contextual features. The problem is important because prediction of communication flow can impact timely sharing of specific information across a wide array of communities. We determine the intent to communicate and communication delay between users based on several contextual features in a social network corresponding to (a) neighborhood context, (b) topic context and (c) recipient context. The intent to communicate and communication delay are modeled as regression problems which are efficiently estimated using Support Vector Regression. We predict the intent and the delay, on an interval of time using past communication data. We have excellent prediction results on a real-world dataset from MySpace.com with an accuracy of 13-16%. We show that the intent to communicate is more significantly influenced by contextual factors compared to the delay.

1. Introduction

In this paper, we develop a computational model for predicting communication flow in large-scale social networks using communication context. The prediction of communication flow is important in determining information propagation through social networks and is useful in applications such as targeted advertising.

There has been prior work on computational models for information diffusion [2,7]. In [2] the authors focus on analyzing the text in blog posts and use an epidemic disease propagation model for determining information diffusion. In [7], the authors present an early adoption based information flow model useful for recommendation systems. There has also been prior work on analysis of emails of software developers [1], to understand the relationship between the email activities and the software roles.

There are several limitations of prior work. Prior research of detecting information flow has been mainly focused on two aspects: (a) implicit assumption of presence of a (virtual) social network through which people can exchange information, and (b) disregard to the local or neighborhood information, semantic content of information, or the points of information generation (sender) and reception (receiver). The model of propagation has been based on static knowledge about people's probability to transmit information. But the nature and degree of propagation are contingent upon the microscopic relationships between people engaged in the process of transmission. Contextual information such as the local network topology of the sender and receiver, relationship of the topic of communication with past communication as well as identity of the recipient, therefore, has not been incorporated. In web based analysis, the flow is estimated from indirect evidence (e.g. a topic appears on a blog several days after it appeared on another blog), not from evidence of direct communication. Thus the effect of contextual factors on communication is not easy to determine.

We describe the interactions between people by two orthogonal yet complementary aspects: *media* and *action*. Every interaction involving two people thus comprises a medium for the propagation of information (e.g. emails, messages, images etc) and an associated action that embodies that interaction (e.g. writing blog posts, adoption of consumer goods etc). The prior work on information diffusion has focused only on the *actions* of the members of the network. But the semantics of the *action* can change under different media contents. Hence analysis of actions for predicting diffusion without consideration of the context in which the information is propagated is limited.

The main contribution of this paper is the development of a contextual framework to predict communication flow between a pair of users. We are motivated by a Physics based wave front metaphor in our understanding of communication flow. This allows us to assume conditional independence in communication from a user's contacts given a topic. We identify three aspects that affect communication on a specific topic: (a) neighborhood context, (b) topic context and (c) recipient context. Neighborhood context refers to the effect of the user's social network on her communication. It is affected by the number of messages by the user's contacts on the topic and the communication in the local neighborhood on the topic. Topic context refers to the effect of the semantics of a user's *past* communication on a topic on her future communication on the same topic. Its features relate to the coherence of messages with respect to each other,

as well as significance of a topic with respect to a user's past communication. Recipient context refers to effect of the recipient identity on the user's intent to communicate. Its features consist of the level of response from the recipient to the user, the topical alignment between the users, and the significance of communication to the recipient with respect to the rest of the communication from the user. The intent to communicate and communication delay are estimated using Support Vector Regression (SVR) on a time interval using their past communication as training data.

We have excellent results on a real world MySpace.com dataset on two sub-problems. We computed the intent to communicate and delay on two topics for a specific user and her network. Secondly, we determined the intent to communicate as well as the communication delay for a single topic, with varying network sizes, averaged over a set of eight contacts of a user. Results show that SVR out performs the baseline technique.

The rest of the paper is organized as follows. Section 2 presents our problem statement. In section 3, we discuss the motivation using a Physics metaphor. Section 4 describes the role of context in communication flow. In sections 5, 6 and 7 we present contextual features. In section 8 we describe a support vector regression method for prediction. Section 9 discusses the MySpace dataset. In section 10, we present the experimental results.

2. Problem Statement

We now present the technical challenges and key issues addressed on this paper. Let Alice and Bob be two users in a social network (Figure 1). Further assume that Alice receives a message on topic Λ. The technical problem addressed in this paper is to predict the communication flow between a pair of users. This is addressed by computing the likelihood that Alice will communicate with Bob on a particular topic and additionally, predict the delay in communication.

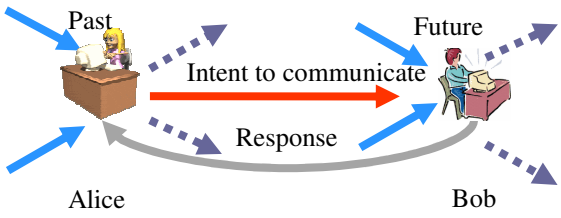


Figure 1: What is the likelihood that Alice communicates with Bob? What is the predicted delay?

The solution to estimate both the likelihood of Alice's communication or the intent to communicate with Bob and the delay lies in determining the message topic and understanding the contextual factors that affect the communication between the two users on this specific topic. There are three contextual factors relating to the sender that are examined in this paper – (a) effect of the local social network of the sender, (b) relationship of message topic to the sender's past communication, and (c) relationship of the sender to the recipient.

3. A Physics metaphor

We are motivated in the analysis of communication flow by a physics based wave metaphor. In the classical wave theory, the phase and magnitude of a wave at a certain point and time in space is the linear superposition of *all* waves from all sources, at the same point in time and space. This superposition can result in constructive (when the phases align) or destructive interference (when the waves are out of phase). Waves are additionally affected by the properties of the medium and exhibit phenomena such as reflection.

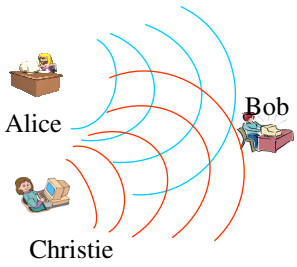


Figure 2: Bob is affected by messages from both Christie and Alice.

We present a simple example to illustrate the role of the metaphor. Let us consider Bob's social network. Alice and Christie are Bob's contacts (Figure 2). We now apply the metaphor to our problem. Bob's communication to his contacts creates a *primary* wave front. When his recipients create messages to further propagate the topic they generate *secondary* wave fronts. Bob will also be the recipient of messages from his contacts – these are reflections or *backscatter*. The multiple communication wave fronts reaching Alice will have constructive/destructive effects on her intent to communicate based on her communication context.

4. The role of context

In this section we discuss communication context and how it is useful to predict communication flow. Communication context [4] is the set of attributes that affect communication between two individuals. Contextual attributes are application dependent and are dynamic. In the application considered in this paper, we identify three aspects that affect communication –

(a) neighborhood context, (b) topic context and (c) recipient context. Let us assume that Alice plans to send a message to her contacts about a topic (e.g. movies) at a certain point of time. Neighborhood context refers to the effect of Alice's contact network on her planned communication. Topic context refers to the effect of Alice's *past* communication about a topic 'movies' on her planned communication about the same topic. Recipient context refers to effect of the recipient identity (i.e. the specific persons with whom Alice communicates) on her intent to communicate. In the next three sections, we shall present our approach to quantifying each of the above contextual attributes.

5. Neighborhood context

Neighborhood context refers to the effect of the user's social network on her communication. There are two network effects of interest – backscatter and susceptibility. We shall use the following example for the following three sections. We have two users Alice (u) and Bob (v) and Bob's contacts (w). Now assume that Alice wants to discuss topic Λ with Bob at a specific time slice t_i. We also denote the count of messages on Λ from u to v as $n_{u\to v}(\Lambda)$.

5.1 Backscatter

Backscatter refers to the fraction of the messages received by u from her contacts that are about a topic Λ. We can reasonably assume that a message sent by u at a certain point of time t_i will be affected by all the messages sent by u's contacts on the same topic Λ that arrived before t_i. We now assume that the number of messages received from a specific contact v_1 is independent of the number of messages from any other contact v_2, given Λ. We further assume that each contact v has the same importance with respect to u. Thus the backscatter due to one contact v in time slice t_i is given by,

$$\theta_{v\to u|u}(\Lambda, t_i) = \sum_{j=1}^{n_{v\to u|u}} \varphi(\Lambda, t_j, t_i), \qquad <1>$$

where, t_j is the time-stamp of the j^{th} message on topic Λ from v to u and $\varphi(\Lambda, t_j, t_i)$ is an indicator function: *1* if t_j lies in time slice t_i and 0 otherwise. Thus, $\theta_{v\to u|u}$ gives the backscatter due to a specific contact v to u and due to an earlier communication from u on the topic Λ. Backscatter due to a specific contact v is proportional to the number of messages sent by v to u. The total backscatter $B_u(\Lambda)$ is given by the sum over all messages received from all L contacts, prior to t_i.

$$B_u(\Lambda) = \frac{1}{NL}\sum_{i=1}^{N}\sum_{v=1}^{L}\theta_{v\to u|u}(\Lambda, t_i), \qquad <2>$$

where N is the total number of time slices prior to t_i.

5.2 Susceptibility

Susceptibility measures whether the social network that u interacts with is interested in the topic that she plans to communicate on. Intuitively, if a network is susceptible to communication on a certain topic, then u is more likely to send a message on topic Λ to her network. Susceptibility is proportional to the number of messages sent by u's contacts to *their contacts*. This is the key difference between susceptibility and backscatter. Making identical assumptions as in the calculation of backscatter, the susceptibility is captured as follows. The susceptibility due to one contact v to her entire social network at time slice t_i is given by,

$$\theta_{v|u}(\Lambda, t_i) = \sum_{w}\sum_{j=1}^{n_{v\to w|u}} \varphi(\Lambda, t_j, t_i), \qquad <3>$$

where, t_j is the time-stamp of the j^{th} message from v to u and $\varphi(\Lambda, t_j, t_i)$ is an indicator function: *1* if t_j lies in time slice t_i and 0 otherwise. Thus $\theta_{v\to w|u}$ gives the susceptibility due to a specific contact v (to her contacts w) due to an earlier communication from u to v on the topic Λ. Susceptibility for a specific topic Λ, for a user u, is just the sum of susceptibilities per contact v, over all past time slices prior to t_i.

$$S_u(\Lambda) = \frac{1}{NL}\sum_{j=1}^{N}\sum_{v=1}^{L}\theta_{v|u}(\Lambda, t_j), \qquad <4>$$

where, N is the number of time slices prior to t_i.

6. Topic context

Topic context refers to the effect of the semantics of a user's *past* communication on the topic Λ on her future communication. We are interested in four measures – (a) message coherence (b) temporal coherence (c) topic relevance and (d) topic quantity.

6.1 Message coherence

Message coherence refers to consistency in message semantics and the semantic relationships of the messages with the current topic Λ (e.g. 'movies').

We use the common sense reasoning toolkit ConceptNet [3] to compute the distance [6] between messages. ConceptNet has several desirable characteristics that distinguish it from the other popular knowledge network – WordNet [5]. First, it expands on pure lexical terms to include higher order compound concepts ("buy food"). The repository represents semantic relations between concepts like *"effect-of"*, *"capable-of"*, *"made-of"*, etc. Finally,

ConceptNet is powerful because it contains practical knowledge – it will make the association that "students are found in a library" whereas WordNet cannot make such associations. Since our work is focused on communication in online social networks, which typically deals with casual conversations, ConceptNet is very useful.

Each message comprises a set of words and is obtained after stop word removal and word stemming on the content of communication. Let $d_c(w_1,w_2)$ denote the ConceptNet distance between two words (concepts) w_1 *and* w_2. Then, the distance between a message m and a topic Λ is given as:

$$d(m,\Lambda) = \max_q \min_k d_c(w_q, w_k), \qquad <5>$$

where, w_q is a word in message m and w_k is a word corresponding to Λ. Given a topic, Wordnet [5] is helpful in determining the synonym set for that topic – this helps us determine the set of words w_k for a topic Λ.

Message coherence $C(\Lambda)$ is then computed as the ratio of $\Omega(\Lambda)$ to $\Omega(\neg\Lambda)$, where $\Omega(\Lambda)$ is the measure of the average similarity of all messages with respect to topic Λ (computed using eq. <5>), and where, $\neg\Lambda$ is the set of antonyms (obtained using WordNet) corresponding to the topic Λ. This ratio has interesting properties: (1) if $C > 1$, then the intent to communicate will be high since messages on Λ are highly coherent; (2) if $C < 1$, then the intent to communicate will be low, since it implies that there is probably a topic in $C(\neg\Lambda)$ which is more coherent than Λ; and (3) if $C \sim 1$, then the effect on the intent to communicate might be considered neutral due to the presence of several topics.

6.2 Temporal coherence

Temporal coherence is defined as the correlation of the time-stamps of the messages on a topic received by a person u. High coherence of messages in a recent past would increase u's intent to communicate and vice versa. Temporal coherence is determined by the mean and variance of the differences in the time stamps of messages received by u in the past referenced from current time t_i. The mean μ_j over a time slice t_j and topic Λ is given by the mean difference of the time-stamps (T) of all the messages in time slice t_j referenced from current time t_i.

$$\mu_j(\Lambda, t_j, t_i) = \sum_{m \in t_j} \left(T(m,\Lambda,t_j) - t_i\right) / n(\Lambda, t_j), \qquad <6>$$

where m is the index of a message of topic Λ in the time slice t_j and where $n(\Lambda,t_j)$ is the number of messages on topic Λ in the time slice t_j. Similarly, the variance σ_j^2 over a time slice t_j and topic Λ is easily computed. Hence for each time slice t_j (duration of one week in our experiments), we can compute mean and variance (μ_j, σ_j^2).

6.3 Topic Relevance and Quantity

Topic relevance for user u on a topic Λ refers to the relationship between *topics in her past communication* to the topic Λ. We can compute topic relevance $\psi_R(u, \Lambda)$ for u on topic Λ by the ratio of the number of messages $n_{u\to v}(\Lambda)$ on Λ sent by u to all her contacts to the total number of messages sent by u to all her contacts on *all topics*.

$$\psi_R(u,\Lambda) = \sum_v n_{u\to v}(\Lambda) \Big/ \sum_\Lambda \sum_v n_{u\to v}(\Lambda) \qquad <7>$$

Topic quantity is the number of topics on which Alice has received messages in the recent past. We determine the number of topics via spectral clustering. We consider messages on the same topic to belong to a specific topic cluster. The number of such topic clusters (k) is determined dynamically in our experiments. The effect of topic quantity $Q_u(\Lambda)$ for a topic Λ on the intent to communicate for user u is inversely related to the number of topic clusters k - $Q_u(\Lambda) = 1/k$. This implies that if u receives messages on many topics (large k), then the intent to communicate will decrease.

7. Recipient context

Recipient context refers to effect of the recipient identity on u's intent to communicate. There are three measures of interest – (a) reciprocity, (b) communication correlation and (c) communication significance.

Reciprocity refers to the ratio of the messages received from the recipient to those sent to the recipient, on the intended communication topic. Reciprocity $r_{v\to u}$ of a user u with respect to v is given by the ratio of the number of messages $n_{v\to u}$ sent by v to u to the number of messages $n_{u\to v}$ sent by u to v on topic Λ. Reciprocity is given as follows:

$$r_{v\to u}(\Lambda) = \frac{n_{v\to u}(\Lambda)}{n_{u\to v}(\Lambda)} \qquad <8>$$

Communication correlation (ρ_{uv}) refers to the topical alignment between a user u and her contact v with whom she wants to communicate. It is computed as a histogram intersection distance over all time slices:

$$\rho_{uv}(\Lambda) = \frac{\sum_i \min(\beta_u(\Lambda, t_i), \overrightarrow{\beta_v(\Lambda, t_i)})}{\sum_i \max(\beta_u(\Lambda, t_i), \overrightarrow{\beta_v(\Lambda, t_i)})}, \quad \text{<9>}$$

where, $\beta_u(\Lambda,t_i)$ refers to the number messages sent by user u at time t_i on topic Λ.

Communication significance refers to the fraction of past messages to the specific contact v on the current communication topic. It is given by the ratio $s_{u\to v}(\Lambda)$ of the number of messages $n_{u\to v}$ from u to v to the number of messages from u to all v on topic Λ.

$$s_{u\to v}(\Lambda) = n_{u\to v}(\Lambda) / \sum_v n_{u\to v}(\Lambda) \quad \text{<10>}$$

We next discuss how our computed features for neighborhood, topic and recipient context are incorporated into a Support Vector Regression based technique. This is used to predict the intent to communicate as well as the average delay in communication.

8. Prediction Framework

The intent to communicate and delay can be modeled as a regression problem where the relationships between the different model parameters can be learnt over time and for specific individuals. To avoid training the time series data every time when we get a new test pair, we use an incremental SVM regression [9] method. The SVR prediction algorithm for intent to communicate is described in Table 1. Estimates of communication delay can be determined using the same procedure as in Table 1.

Table 1. Algorithm for SVR prediction.

Input: $(x_i, y_i) \in X \times \mathbb{R}$, $X = \mathbb{R}^d$ pairs of training data.

a. Predicted intent x_i, $i= 1, 2, \ldots, N$ where N is the number of time slices over past communication for two *specific* users u and v on topic Λ. It consists of the contextual feature vectors, $x_i= \{S_u(\Lambda), B_u(\Lambda), C, (\mu_i, \sigma^2_i), \psi_R(u, \Lambda), Q_u(\Lambda), r_{v\to u}(\Lambda), \rho_{uv}(\Lambda), s_{v\to u}(\Lambda)\}$.

b. Actual communication (based on frequency count of messages sent by u to v) y_i $i= 1, 2, \ldots, N$ where N is the number of time slices over past communication for users u and v on a topic Λ.

Procedure:

a. The SVM regression function $f(x)$ is trained on $\{(x_1, y_1), \ldots, (x_N, y_N)\}$. It is tested on the incremental sample (x_{N+1}, y_{N+1}).

b. The training set is then augmented with the new samples (x_{N+1}, y_{N+1}) and a new regressor is learnt using this training set.

c. Repeat a-c until there are no more samples.

Output: Error in prediction, E.

a. Use $f(x)$ with training set $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ and determine predicted $\hat{y}_{N+1}$.

b. Determine the actual communication intent y_{N+1}.

c. Compute error as, $E= (y_{N+1} - \hat{y}_{N+1}) / y_{N+1}$.

The use of the SVM based regression algorithm allows us to incrementally predict the intent to communicate and the communication delay for a user u with a specific contact v.

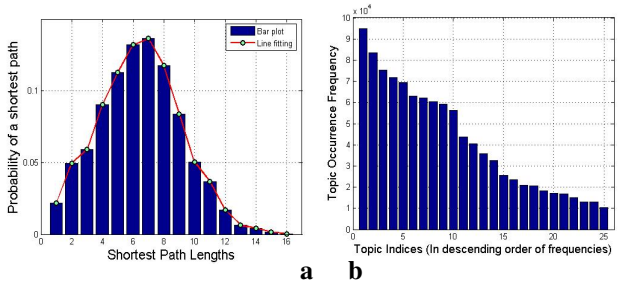
9. The MySpace dataset

MySpace is the world's largest social networking site with over 108 million users. The dataset used for our experiments comprises approximately 20,000 users who have exchanged about 1,425,010 messages in the time snapshot from September 2005 to April 2007. The crawling process was seeded from one of Tom's (super-user of MySpace who is a contact of all the users) top eight friends. A depth first strategy was adopted to continually crawl the friends of a user who have sent messages to the user in the said time period. The process was continued till we reached the third level friend of each user (friends-of-friends). This strategy was adopted to ensure that the data we are dealing with exhibits sufficient traces of communication among the crawled users as well as the network of users was sufficiently cohesive in relationship pertaining to communications.

We describe the network topology of the crawled dataset using three standard metrics: average shortest path length, degree distribution and clustering coefficient. Our analysis shows that the average shortest path length μ is approximately 5.952 (Figure 3(a)). The degree distribution is long-tailed and follows a power law distribution $P(k)\sim k^{-\gamma}$ (γ is a network coefficient) with γ= 2.01. Finally the clustering coefficient, defined as the probability that friends of a person will mutually be friends too, was determined to be 0.79. These measures are consistent with statistics of other social network datasets [8] which follow a topology akin to scale-free networks and observe the 'small-world phenomenon'.

We now discuss how each message is assigned a topic. This is done using a simple aggregation algorithm that exploits the tree structure of using WordNet [5]. In WordNet, each word belongs to a synonym set

(synset), representing a unique lexical concept. For each word in the message, we determine the synset to which belongs. Using WordNet, we also determine the third-level generalization for each synset. These typically are abstractions ('entity, physical, object' is the third level generalization for the concept 'car') and we refer to them as topics in this paper. Two lower level synsets are similar, if they share the same topic. We assign the topic to a message that covers the largest number of message synsets. In this research we examine the 25 most frequently occurring topics concerning about 15,000 users and 1,140,000 messages exchanged between them. The topic histogram is shown in Figure 3(b). The two most frequently occurring topics have been used for our experiments.



a **b**

Figure 3(a): Average Path Length Distribution. (b) Topic Histogram.

10. Experimental results

In this section we discuss the experimental results conducted based on the model. We first discuss the baseline techniques used for evaluating the model. We then describe temporal dynamics of the features followed by the results of the predicted intent to communicate and the delay. Finally we discuss the evaluation of the individual contextual features.

10.1 Baseline techniques

The baseline technique for predicting the intent to communicate is computed using the prior probability of communication on topic Λ. This probability is proportional to the frequency count of messages exchanged between the users u and v in the past on Λ. The predicted intent I is given by the ratio of the number of messages n sent by u to v on topic Λ to the total number of messages on all Λ sent by u to v in the past, described as below:

$$I_{u\to v}(\Lambda) = n_{u\to v}(\Lambda) \Big/ \sum_{\Lambda} n_{u\to v}(\Lambda) \qquad <11>$$

The baseline technique of prediction of delay is based on determining the correspondences between two messages between the users and then computing the mean delay by examining the message time stamps.

For determining the semantic correspondence of messages we use the ConceptNet distance (ref. eq. <5>) – for each message, we assume that the nearest message (in terms of time) whose distance is below a threshold is the corresponding message. We acknowledge that the message correspondence issue will benefit through a linguistic analysis of messages – that is beyond the scope of this paper. Then the mean delay between two contacts u and v on topic Λ is the mean delay between all pairs of corresponding messages on the same topic. The predicted delay $D_{uv}(t_{m+1}, \Lambda)$ for the next time slice t_{m+1} is computed at time t_m as the mean delay across all previous time slices till t_m:

$$\hat{D}_{uv}(t_{m+1}, \Lambda) = \frac{1}{m}\sum_{i=1}^{m} D_{uv}(\Lambda, t_i). \qquad <12>$$

10.2 Feature temporal dynamics

In this section, we describe some simple visualization showing the temporal dynamics of the different contextual features. We consider a single user's local social network comprising eight people averaged on two topics A and B. The dynamics are shown over duration of four weeks. Description of the topics with examples is shown in Table 2.

Table 2. Topic abstractions used for experiments.

Topic	Abstraction	Example message
Topic A	'entity, physical, object'	"Hey Julia, were you able to find your *car* in the *parking lot*? It has been hell of a day today with the *car* search…"
Topic B	'person, someone, human'	"The party went off fine. We were eagerly waiting for *Annie* to come though. *She* would really make a difference!"

Figure 4 (a-d) shows the visualization of the contextual features. Each arm shows the corresponding entity between the selected user and one of eight contacts, also each column is a week's representation. The features are: communication correlation (high value proportional to edge length), communication significance (high value proportional to area of a sector in pie), reciprocity (proportional to share of length in an edge) and backscatter (high value proportional darker shade of the circular area). The final intent to communicate and the delay are shown in (e-f). The

visualization of the features shows how the different features interact to yield high or low values of intent (proportional edge thickness) and delay (proportional to edge length). It also shows that the values of the features change temporally resulting in varying intent and delay with respect to a particular pair of users.

10.3 Specific single-node network

In this section, we present the results of the estimates for intent to communicate and the delay concerning the topics A and B. This is done for a single user's local social network comprising eight people, over duration of five weeks. We also show the average error in prediction across all 25 topics for five weeks.

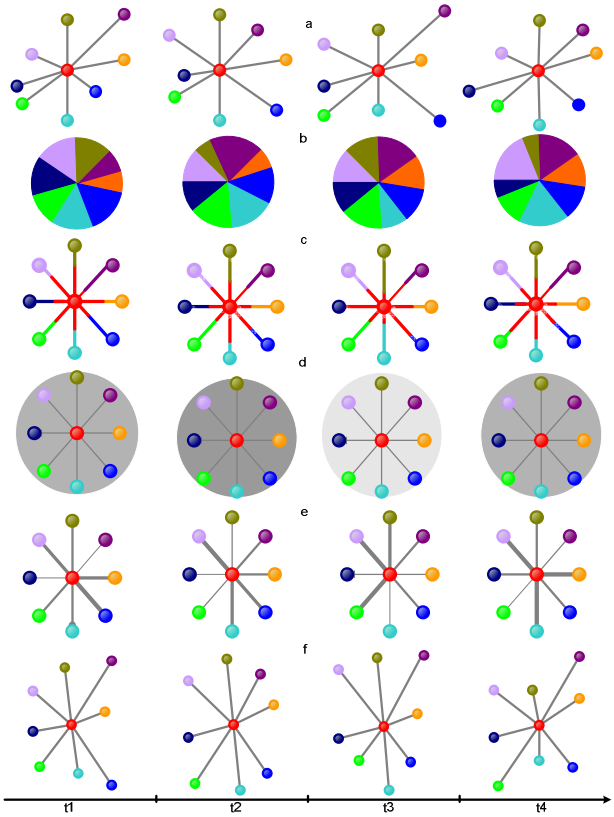


Figure 4(a-f): Communication Correlation, Communication Significance, Reciprocity, Back Scatter, Predicted Intent, Delay.

Figure 5(a) and (b) show the mean intent to communicate for *u* for *each* of her eight contacts for the two topics A and B respectively. The values are averaged over all five weeks. The figures show three measures per contact – (a) the actual communication, (b) baseline communication using eq. <11> and (c) our SVR based method (ref. Table 1). The figures reveal that errors over the five weeks in the intent to communicate for the SVR prediction is between 10~15% (mean error: Topic A: 12.83%, Topic B: 13.46%), while the error for the baseline technique is between 40~65% (mean error: Topic A: 49.21%, Topic B: 45.49%). Our explanation for this discrepancy is as follows. The communication intent depends on a wide variety of contextual factors (neighborhood, topic, and recipient) and not just on prior probability of communication on that topic. We believe that our approach captures these important contextual factors, yielding effective results.

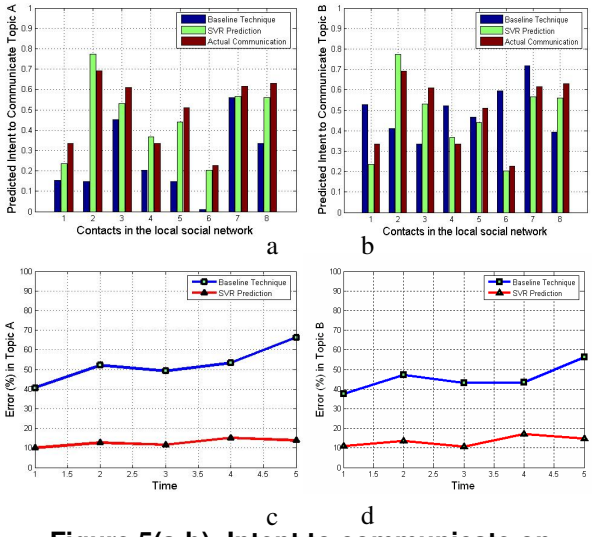


Figure 5(a-b): Intent to communicate on topics A and B. (c-d): Error in prediction of intent to communicate for topics A and B over five weeks.

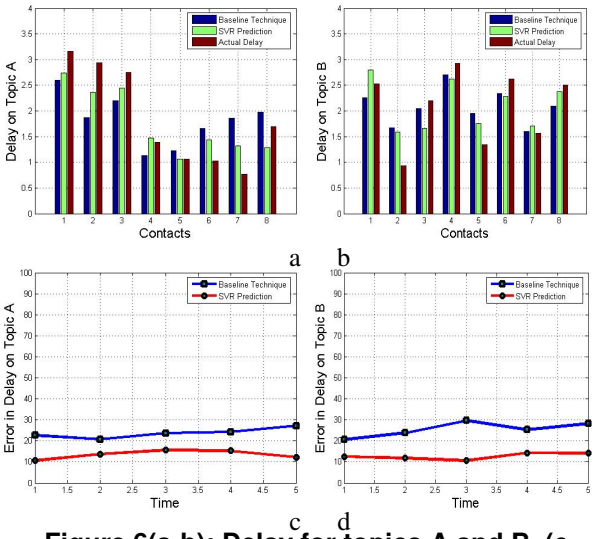
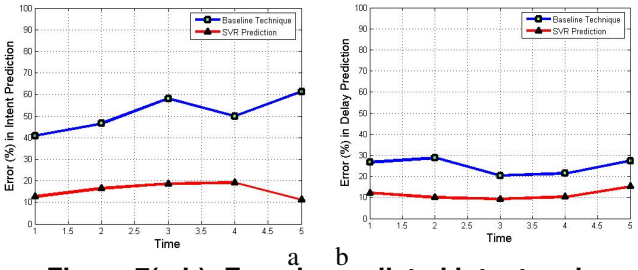


Figure 6(a-b): Delay for topics A and B. (c-d): Error in prediction of delay for topics A and B over five weeks.

Figure 6(a-b) shows the mean predicted delay for *u* for each of her eight contacts, averaged over five weeks for topics A and B respectively. The figure again shows three numbers as before, per contact.

The figures (Figure 6 (a-b)) reveal that errors over the five weeks in the delay estimate for the SVR prediction is between 10~15% (mean error: Topic A: 13.92%, Topic B: 15.04%), while the error for the baseline technique is between 20~35% (mean error: Topic A: 24.28%, Topic B: 28.48%). Interestingly, the baseline delay estimate (prior probability) works reasonably well although not as well as the SVR technique. We conjecture that delay for a single person may be strongly influenced by factors other than the social network interaction (e.g. they may be habitual). The results of the intent and delay hold equally good for the average across 25 topics as well (Figure 7(a-b)).



a b

Figure 7(a-b): Error in predicted intent and delay averaged across 25 topics.

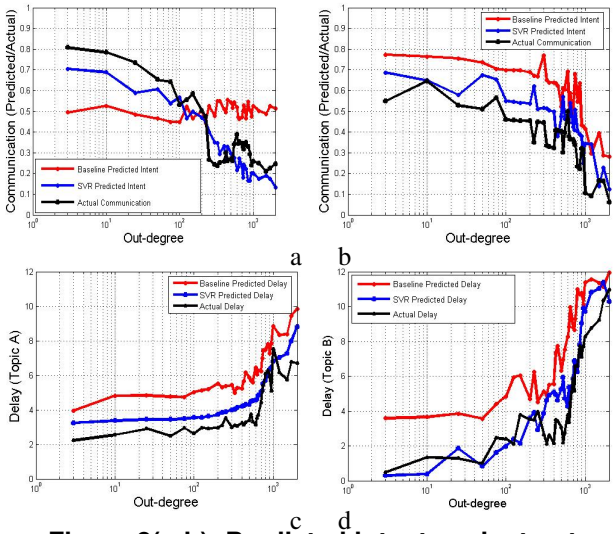
10.4 Network scaling properties

In this section, we present the results of the predicted intent to communicate and delay for two topics (A and B), for a single time slice, but with varying social network sizes. The goal is to understand the role of the size of the social network on the prediction results.

We created a set of networks by sampling the MySpace dataset. We used an exponential function: $f(n)= exp(n/k)$, where k= 4.6 and n= 1, 2, 3, 4, …, 35 to choose networks with node out-degree values $f(n)$. We selected the top three users corresponding to each $f(n)$ based on high message density (number of messages exchanged by the user with her social network) using the MySpace dataset. The intent to communicate on topic A and B for each of these three users with their individual social networks is determined. The mean intent to communicate per network size n is then the mean of the intent to communicate for the three users (whose network size is n), with their networks.

We observe from Figure 8(a-b) that the predicted SVR intent follows a gradual decay as the out-degree increases. The SVR prediction outperforms the baseline technique with a mean error of 18-20% compared to the actual communication. Note however that the SVR technique follows the actual communication curve, while the baseline technique is fairly stable. The overall decrease may be explained as follows. With an increase in network size, the user may be in regular correspondence with only a small fraction of the network. Since we calculate the average over all contacts, this leads to an overall decrease in the intent.

The analysis of delay prediction for the same topics A and B based on the variation of network size is revealing. We observe from Figure 8(c-d) that with increase in out-degree, the mean delay increases with increase in network size. We believe that this is reflective of the fact that users may only correspond regularly with a small fraction of their network – since we take the average over all users, this is influenced by a majority with whom the user is not in active communication. Interestingly, the baseline technique again performs well, indicating that the delay may be due to intrinsic factors (e.g. habitual) and less affected by the contextual factors. The errors in delay for the SVR case are between 12% and 20%. The errors for the baseline case are between 23% and 35%.



a b c d

Figure 8(a-b): Predicted intent against out-degree for Topics A and B. (c-d): Predicted delay against out-degree for Topics A and B.

10.5 Evaluation of features

In this section, we discuss the evaluation of the several contextual features used in our prediction model. The results span across the topics A and B and over a single person's social network of eight contacts for three consecutive time slices (three weeks).

For evaluation of each individual feature, we adopted the L-O-O (or Leave-One-Out) procedure. We

determine the error in prediction of the intent and delay leaving one feature out at a time. Figure 9(a-b) gives the errors in prediction of the intent to communicate and (c-d) gives those for the predicted delay. In each set of the nine bars, each bar corresponds to the error in prediction when a particular feature (indicated in the legend) is left out. From analysis of the errors, we notice that five features: susceptibility, back scatter, message coherence, communication significance and reciprocity, when left out, negatively affect the prediction, implying their significance.

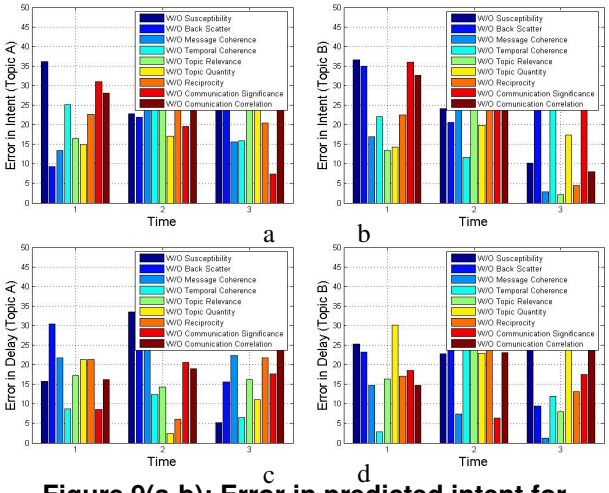


Figure 9(a-b): Error in predicted intent for topics A and B. (c-d): Error in predicted delay for topics A and B.

The evaluation of the features leads interesting insights into the importance of context in predicting communication flow. We observe that a person's neighboring social network indeed effects whether or not she will engage herself in a particular communication quickly. It is also apparent that a person's desire to respond to a communication relies greatly upon the past feedbacks she got from her contacts. The results also emphasize that communication flow is a function of variability in the topic of discussion. Again, for a more elaborate prediction methodology, we might train the SVM Regressor based on the L-O-O procedure to learn the weights corresponding to each feature. However, this is beyond the scope of current work.

11. Conclusion and future work

We have developed a novel framework to predict communication flow in a large scale social network based on communication context. We used a physics based metaphor to motivate our analysis. We identified three aspects that affect communication on a specific topic: (a) neighborhood context, (b) topic context and (c) recipient context. The intent to communicate and communication delay were estimated using Support Vector Regression over a set of contextual features. We have excellent results on a real world MySpace.com dataset on two different scenarios – for a single user as well as over networks of different sizes. Our results show that SVR out performs the baseline technique, with significantly smaller error on both problems. Interestingly while the intent to communicate is strongly affected by the contextual factors, the delay is less affected suggesting that factors external to the social network may be responsible.

There are several interesting directions to future work: (a) Comparison against a standardized flow model e.g. epidemic disease propagation model (b) Prediction, given a pair of users who are separated by *n* different people in the social network (c) Contextual correlation or coupling between contextual features and (d) Temporal evolution of communication context using a partially observable Markov decision process (POMDP). People engaged in communication can be assumed to exhibit variable communicative behavior under partially observable finite states (e.g. locations, time zones etc).

12. References

[1] C. BIRD, A. GOURLEY, *et al.* (2006). *Mining email social networks*, Proceedings of the 2006 international workshop on Mining software repositories, 137-143, Shanghai, China.

[2] D. GRUHL, R. GUHA, *et al.* (2004). *Information Diffusion through Blogspace*, Proceedings of the 13th international conference on World Wide Web,

[3] H. INO, M. KUDO, *et al.* (2005). *Partitioning of Web Graphs by Community Topology*, Proceedings of the 14th international conference on World Wide Web, 661-669, Chiba, Japan.

[4] A. MANI and H. SUNDARAM (2006). *Modeling User Context with Applications to Media Retrieval.* to appear in Multimedia Systems Journal, Summer 2006.

[5] G. A. MILLER, R.BECKWITH, *et al.* (1993). *Introduction to WordNet : An on-Line Lexical Database.* International Journal of Lexicography **3**(4): 235-244.

[6] B. SHEVADE, H. SUNDARAM, *et al.* (2007). *Modeling Personal and Social Network Context for Event Annotation in Images*, Proc. Joint Conf. on Digital Libraries 2007, Jun. 2007, Vancouver, Canada.

[7] X. SONG, B. L. TSENG, *et al.* (2006). *Personalized recommendation driven by information flow*, Proc. 29th ACM SIGIR Conference, ACM Press, 509-516, Seattle, Washington, USA.

[8] D. J. WATTS and S. H. STROGATZ (1998). *Collective dynamics of 'small-world' networks.* Nature **393**(6684): 440-442.

[9] F. ZHANG (2000). *An approach to incremental SVM learning algorithm*, Proceedings of the 12th IEEE international Conference on Tools with Artificial intelligence, IEEE Computer Society, November 13 - 15, 2000, Washington DC.