2009 International Conference on Computational Science and Engineering

SOCIAL SYNCHRONY: PREDICTING MIMICRY OF USER ACTIONS IN ONLINE SOCIAL MEDIA

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

ABSTRACT

We propose a computational framework to predict synchrony of action in online social media. Synchrony is a temporal social network phenomenon in which a large number of users are observed to mimic a certain action over a period of time with sustained participation from early users. Understanding social synchrony can be helpful in identifying suitable time periods of viral marketing. Our method consists of two parts – the learning framework and the evolution framework. In the learning framework, we develop a DBN based representation that includes an understanding of user context to predict the probability of user actions over a set of time slices into the future. In the evolution framework, we evolve the social network and the user models over a set of future time slices to predict social synchrony. Extensive experiments on a large dataset crawled from the popular social media site Digg (comprising ~7M diggs) show that our model yields low error (15.2±4.3%) *in predicting user actions during periods with and without synchrony. Comparison with baseline methods indicates that our method shows significant improvement in predicting user actions.*

1. INTRODUCTION

This paper presents a framework for predicting social *synchrony* in online social media. By synchrony, we mean the tendency of a large group of people to perform similar actions in unison, in response to a contextual trigger. Consider the familiar observation from a performance in an auditorium. When a small set of individuals starts to clap, the rest of the audience follows. This lasts for a short period of time, till the claps die. In Nature, biological oscillators, including fireflies are often observed to fire light at the same time, triggered by a few of them. This phenomenon of oscillating light continues for a certain period of time. Note in both examples, the population moves to a state of *sync* with respect to a certain action (clapping and light oscillation), in response to trigger by a small set of participants.

Do social media websites including Facebook and Digg, exhibit this temporal property of sync? These sites provide a variety of user affordances including observing their friend behavior, commenting, and posting / sharing media content. For example on Facebook, certain types of behavior can be seen to be performed in near unison– e.g. taking of online quizzes, using certain applications to send message to one's contacts; while on Digg, this can include liking or sharing stories around a particular topic. There is typically a contextual trigger (typically one's contacts or a set of users the user is familiar with), that prompts similar behavior from the user. An important motivation for this paper is to identify the conditions under which large number of users in a social network exhibit *synchrony* in mimicking a certain action.

The synchrony problem is related to, but distinct from, the problem of network cascades [9,11]. Unlike a cascade, in the examples discussed above, the original set of seed users who trigger the behavior continue to participate in performing the same set of actions. Cascades do not include the idea of performance in unison, or the idea of sustained participation. Like cascades, synchrony must attract new users to act along with the existing users.

Understanding social synchrony can be of utility in several diverse domains. Corporations might be interested to know what could be the suitable time periods to market a particular product. Predicting time periods involving high user participation in a network can also be useful in resource allocation and management. It can also help us understand the responses of user groups to certain types of events.

Related Work. Although social synchrony has not been studied in prior work, there has been work on related ideas in peer disciplines which have dealt with the response of a social network to dynamic information – epidemiology and information flow [2,6,9], social cascades [4,10,11], social correlation [3] and social recommendations [5,8].

Gruhl et al [6] develop a topic propagation framework and use a model of infectious diseases to study information diffusion. Adar and Adamic in [2] study the phenomenon of information epidemics in the blogosphere based on the idea of propagation of memes. Leskovec et al in [9] use the property of sub-modularity to develop an optimization algorithm that detects points of origin of outbreaks in networks. In [11] Watts analyzes the conditions under which global information cascades occur, based on a simple threshold idea of changing user states. Leskovec et al in [10] attempt to understand how realistic cascades emerge in blogs and study information propagation in the blogosphere.

The prior work address issues such as information flow or emergence of social cascades well. However, they assume that user participation in an action occurs at a single point in time without continued participation, and without the idea that *addition* of new users affects the dynamics of the phenomena over time. In online social media, these assumptions are not always true – for example, a user on Digg can continually 'digg' news stories on a topic over a period of time, and this continued participation can impact other new users in the network to participate as well. To the best of our knowledge, this paper is one of the first attempts to study such temporal properties characterizing synchrony in social networks.

978-0-7695-3823-5/09 $26.00 © 2009 IEEE
DOI 10.1109/CSE.2009.439

Our Approach. There are two key contributions in this paper:

1. First, we propose an operational definition of synchrony of user actions in online social media. Our definition incorporates the following ideas – presence of a specific topic, an agreed upon action, a seed set who triggers an action and high frequencies of continuing old users as well as new users over a period of time.
2. Second, we develop a computational model to predict emergence of synchrony over a period of time. It uses a learning framework and an evolution framework. In the former, a dynamic Bayesian representation of a user based on latent states and user context is developed to predict her probability of a specific action at a certain point in time. In the latter the social network of users is evolved to account for varying network sizes and user models over time.

We have conducted extensive experiments on a very large dataset crawled from the social media site Digg [1] (comprising ~7M diggs). The results show that our model yields low error (15.2±4.3%) in predicting user actions (or diggs) during time periods with and without synchrony. Comparison with baseline techniques indicates that our method shows significant improvement in prediction of user actions in the range 18-56%.

The rest of the paper is organized as follows. In section 2, we present our notion of social synchrony. Section 3 presents our problem definition. In sections 3 and 4, we present our computational framework. In section 5, we discuss our experimental results. We conclude in section 6 with our major contributions.

2. WHAT IS SOCIAL SYNCHRONY?

We now examine a popular real world social network Digg [1]. Users on Digg can *vote* on news stories posted on an external website (popularly known as 'digging'). They can also express their disapproval over items by 'burying' them.

As an example, consider the pool of stories on two different topics on Digg – 'Olympics' and 'Celebrity'. Figure 1 shows the number of new users (i.e. the users who digg a story on the particular topic for the first time) on each day over the month of September 2008 and continuing old users (i.e. the users who have dugg a story on the topic before) across each day for each of the two topics.

We observe that the topic 'Olympics' is characterized by a time period of large number of new users (Sept 3-Sept 13). During this period, we also observe considerable numbers of continuing old users. On the other hand, in case of the topic 'Celebrity' we observe that the number of continuing users is rapidly decreasing over time and the rate of new users joining the existing user set is nearly constant. Hence in the case of 'Olympics', the rate of new users getting involved in the action of digging is high, as well as several old users continue to perform the action of digging stories on the topic. We conjecture that the topic 'Olympics', unlike 'Celebrity' exhibits social synchrony (over the action of digging) between Sept 3 and Sept 13, in a manner similar to the earlier examples of clapping and light oscillation in the previous section.

These examples indicate that synchrony given a topic would involve two properties: (a) sustained participation from old users over time in perform a specific action – we call this property *sustainability* α, and (b) attraction of large number of new users over time to perform the same action – we call this property as the *rate of attraction* β. The measure of sustainability α is therefore given by the ratio of the number of common users across consecutive time slices, to the total number of users in the previous time slice. Similarly, the rate of attraction β is further given by the ratio of the number of new users at a time slice, to the total number of unique users across all time slices.

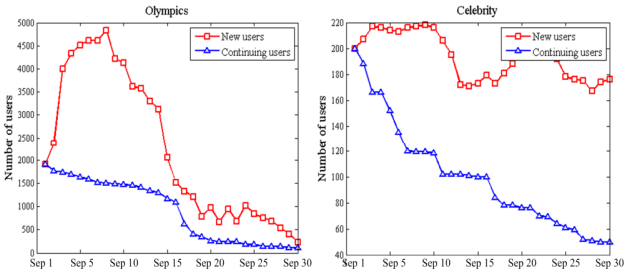

Figure 1: Dynamics of the size of user set on two topics from Digg – 'Olympics' and 'Celebrity' (September 1-30, 2008). Topic 'Olympics' is observed to exhibit synchrony where old users continue to be involved in the action of digging stories, as well as large number of new users join in the course of time (Sept 3-Sept 13).

Definition. Social synchrony is therefore a temporal phenomenon occurring in social networks which is characterized by (a) a certain topic (including a meaningful theme), (b) an agreed upon action(s) that the users in the network can perform with respect to the topic, (c) a set of seed users who are involved in performing the action at a certain point in time, and (d) large numbers of continuing old users as well as new users getting involved in the action over a period of time in the future, following the actions of the seed set (i.e. sustainability α~1 and rate of attraction β››1).

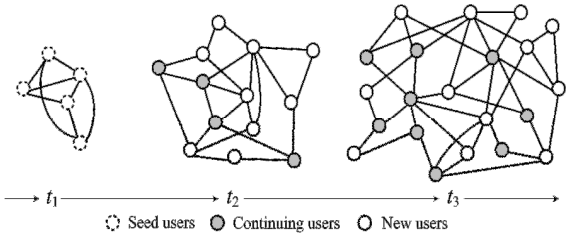

Figure 2: Conceptual representation of synchrony in a social network over three time slices: t_1, t_2 and t_3. Synchrony is likely to involve (a) seed users, (b) sustained participation from old users (shaded circles) as well as attract new users (white circles) to perform a specific action over a period of time.

We explain the idea of social synchrony in the conceptual representation in Figure 2. Starting with a small set of seed users, the figure shows how large numbers of new users join the initial set in performing a certain action (white circles). It

further shows sustained participation from the old users (shaded circles) over a period of three time intervals.

3. PROBLEM DEFINITION

Now we present our data model, followed by the problem statement and finally, the key challenges.

Data Model. Our data model is based the popular news-sharing social media, Digg. News stories on Digg are organized into a two-level Digg-defined content taxonomy – the higher level theme being called a 'container'; while each container further comprising a number of 'topics'. To study social synchrony on Digg, we focus on stories at the granularity of a topic. Each user on Digg can have a set of contacts, whose digging actions on stories of different topics can be viewed in the user's Digg profile through feeds.

It is important to note that a user can digg a story under two conditions. First, the user can be present on the Digg website while digging an item. This can be described to be a *socially aware* state because her action of digging a story in this case is likely to be impacted by the actions of other Digg users. Second, she can digg it from the source website of the news story itself (e.g. digg a news story on the NY Times website), which we call the *socially unaware* state. In this paper we will use the symbol '1' to represent the socially aware state, and '0' to represent the socially unaware state (ref. Figure 3). Note the state of the user is hidden – we have no knowledge about the condition under which a particular item was dugg by a user.

Besides the action of digging, users can further engage in communication by posting comments on an already dugg story (i.e. a story that has been voted upon earlier by other users), or involve in discussion with other users via replies to existing comments.

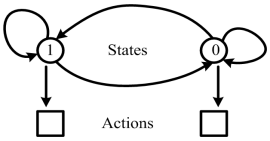

Figure 3: Representation of the two conditions (via two latent states) under which a user can perform a specific action – socially aware (e.g. digging on the Digg website), denoted by '1' and socially unaware states (e.g. digging from an external site like the NY Times), denoted by '0'.

Although we focused on the social media Digg in this paper, our framework can be easily extended to other social networks with evidence of user actions and user communication. Several other social media sites like Flickr, del.icio.us, YouTube, Yahoo! Buzz, Reddit, StumpleUpon, and social networks like Facebook and MySpace also allow users to execute specific actions both on and off their website – hence the notion of the two conditions under which users can perform a specific action is also extensible.

Problem Statement. We now frame our problem statement as follows. Given:

1. a topic ζ and a social network $G(V,E)$ where V is the set of users and E is the set of edges between them. An edge $e_{uv} \in E$ exists if u and v are contacts of each other;
2. a set of actions Ψ: $\{a_1, a_2, \ldots\}$ and a set of communication types τ: $\{x_1, x_2, \ldots\}$ that can be performed by each user $u \in V$ over the topic;
3. a seed set of users U_0 who perform actions, $a_i \in \Psi$ with respect to the topic ζ at a certain time slice t_0; and
4. a history of actions a_i and communication x_i (comments and replies among users in V) on the same topic ζ and over a period of T time slices (prior to t_0) corresponding to each user, $u \in V$,

we are interested in predicting whether the network G will exhibit *social synchrony* (operationally defined in section 2) over M future time slices after t_0 (t_1-t_M) on the same topic.

Key Challenges. Predicting an emergent social synchrony with respect to a topic and over a set of M future time slices involves the primary challenge of determining the temporal evolution of the social network G with respect to the predicted actions performed by its users (Figure 2). This, in turn, involves the following sub-challenges:

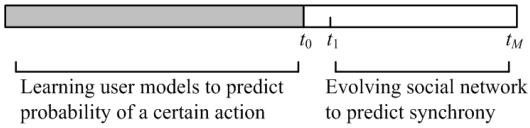

Figure 4: Illustration of the learning and evolution problems in prediction of synchrony.

1. *Learning* – for each user in the social network, we need to predict her probability of actions at each future time slice (Figure 4). The probability estimate of user action should take into account the condition under which the action is performed (socially aware / unaware state) as well as how her context affects it. The user context should further incorporate the effect of the actions of other users in the network in the previous time slice, degree of coupling with the seed set with respect to the specific action, as well as the user's own history of actions and communication.
2. *Evolution* – recall, synchrony in a social network (a) is likely to involve sustained participation from old users as well as attract new users to perform a certain action; and (b) persists over a period of time. Hence, we need to determine the network size at each of the M future time slices, choose the set of users in each of the evolved networks, and finally determine estimates of the probability of actions of all users in the evolved networks at each future time slice. These probability estimates would indicate the presence (or absence) of an emergent synchrony in the future for the particular topic (Figure 4).

In the following two sections, we discuss our framework that addresses each of the above challenges.

4. THE LEARNING FRAMEWORK

In this section we present a framework for predicting the probability of actions of each user in the social network, given a topic, at each of the M future time slices. We need to solve four related sub-problems: (a) represent the user actions as a dynamic Bayesian network (DBN), (b) estimate the user

context, (c) predict probabilities of the user states, and (d) predict probabilities of her actions.

4.1 DBN Representation

A user's intent to perform an action (on a certain topic) can be triggered due to her being either *socially aware* or *socially unaware* of the actions in her local social network i.e. her immediate neighbors (section 3). Note each of the two states, in turn, would be affected by the user context (e.g. actions of the neighboring contacts, coupling with seed users and / or the user's communication over the topic).

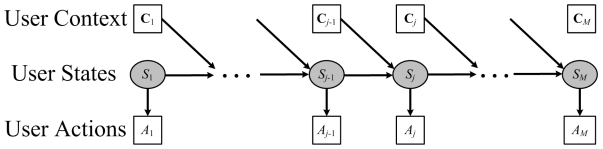

Figure 5: Dynamic Bayesian Network representation of user actions using user context and states.

We represent the temporal dependencies among all these variables – user actions, states and contextual attributes via a dynamic Bayesian network shown in Figure 5. Assuming first order Markov property, the DBN yields the following time-based inference equation for determining the probability of user action at a time slice t_j ($1 \le j \le M$):

$$P\left(A_{u,j} \mid \vec{A}_{u,j-1}, \vec{\mathbf{C}}_{u,j-1}\right)$$
$$= \sum_{S_{u,j}} P\left(A_{u,j} \mid S_{u,j}, \vec{A}_{u,j-1}, \vec{\mathbf{C}}_{u,j-1}\right) \cdot P\left(S_{u,j} \mid \vec{A}_{u,j-1}, \vec{\mathbf{C}}_{u,j-1}\right) \quad <1>$$
$$= \sum_{S_{u,j}} P\left(A_{u,j} \mid S_{u,j}\right) \cdot P\left(S_{u,j} \mid S_{u,j-1}, \mathbf{C}_{u,j-1}\right),$$

where $A_{u,j}$ is the action performed by user u at time t_j, $S_{u,j}$ is the state of u at time t_j, $\mathbf{C}_{u,j-1}$ is the context of user u at the previous time slice, the vectors $\vec{A}_{u,j-1}$ and $\vec{\mathbf{C}}_{u,j-1}$ represent $A_{u,1}$, $A_{u,1}$,…, $A_{u,j-1}$ and $\mathbf{C}_{u,1}$, $\mathbf{C}_{u,2}$, …, $\mathbf{C}_{u,j-1}$ respectively. We show how each of the probabilities are estimated in the following subsections. First we present how we estimate user context.

4.2 User Context

We discuss the different attributes of the user context that are associated with the two latent user states. Let us denote the contextual attributes at time slice t_j corresponding to the socially aware state by $\mathbf{C}^1_{u,j}$. The different attributes in $\mathbf{C}^1_{u,j}$ are as follows:

1. A user is in the socially aware state would be affected by the actions of her neighborhood, i.e. the set of immediate contacts (Figure 6(a)). Hence the first contextual attribute is the mean probability of actions over user u's neighborhood in the previous time slice. This is denoted as $\eta_{u,j-1}$.
2. A user u in the socially aware state would also be affected by the coupling $\omega_{u,j-1}$ with the seed set of users U_0 in the previous time slice (Figure 6(b)). It is given by the ratio of how many times u followed v's actions on a topic in the previous time slice, to the total frequency of u's actions at that time slice.
3. The prior probability of u's own actions $p(A_{u,j-1})$ in the previous time slice; and
4. The user u's intrinsic interest on the particular topic in the previous time slice – her probability of communication. This is estimated via the frequency of commenting, $p(\kappa_{u,j-1})$ and replying, $p(R_{u,j-1})$.

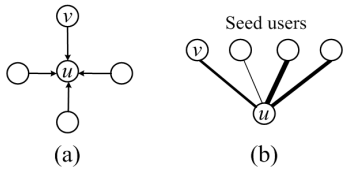

Figure 6: (a) Effect of a user's neighborhood (i.e. contacts) on her actions, shown by the directed edges. (b) Coupling of a user's actions with the seed users. The degree of coupling is represented using the thickness of an edge. Note the semantics of the two edges are different – a seed user in (b) may or may not be a contact of *u* in (a).

The attributes in $\mathbf{C}^0_{u,j}$ associated with the socially unaware state are also determined in a similar manner. However since the user does not have knowledge about her neighborhood or the seed users in this state, hence $\eta_{u,j-1}$ and $\omega_{u,j-1}$ are assumed to be equal to zero.

4.3 Estimating User States

We now present our method of estimating the probability of the user states.

Let us assume that the probability of a certain user state at t_j given the previous state and the user context at t_{j-1} ($P(S_{u,j} \mid S_{u,j-1}, \mathbf{C}_{u,j-1})$) is given by a probability density function whose parameters are unknown to us. To estimate this probability for a specific set of contextual attributes in $\mathbf{C}_{u,j-1}$ our goal is therefore to determine the model parameters of the pdf that describes this conditional probability. In this paper we assume that the user states $S_{u,j}$ at t_j have a multinomial density [12] over the contextual attributes in $\mathbf{C}_{u,j-1}$ (with parameter $\varphi_{u,j-1}$), along with a conjugate Dirichlet prior [12] over the previous state $S_{u,j-1}$ (with parameter $\gamma_{u,j}$). This is given by,

$$P\left(S_{u,j} \mid S_{u,j-1}, \mathbf{C}_{u,j-1}\right) \propto P\left(\mathbf{C}_{u,j-1} \mid S_{u,j}\right) P\left(S_{u,j} \mid S_{u,j-1}\right)$$

$$\text{where } P\left(\mathbf{C}_{u,j-1} \mid S_{u,j}\right) = \frac{\sum_{ik} \mathbf{C}_{u,j-1;ik}!}{\prod_{ik} \mathbf{C}_{u,j-1;ik}!} \prod_{ik} \varphi_{u,j-1;ik}^{\mathbf{C}_{u,j-1;ik}} \quad <2>$$

$$\text{and} \quad P\left(S_{u,j} \mid S_{u,j-1}\right) = \frac{1}{B\left(\gamma_{u,j}\right)} \prod_{il} S_{u.j}^{S_{u.j-1;il}},$$

and $B(\gamma_{u,j})$ is a Beta function with the parameter $\gamma_{u,j}$. Based on eqn. <2>, let $\lambda_{u,j-1}$ denote the parameters corresponding to the pdf of $P(S_{u,j} \mid S_{u,j-1}, \mathbf{C}_{u,j-1})$. Our goal is to maximize the likelihood of $\lambda_{u,j-1}$ based on the temporal relationships between user states and context in the DBN in Figure 5:

$$\lambda^* = \arg\max_{C_{u,j-1}} \log P\left(S_{u,j} \mid \mathbf{C}_{u,j-1}, S_{u,j-1}\right). \quad <3>$$

Assuming independence between the user state and the user context at the same time slice, and using first order Markov

property in our DBN, it is sufficient to maximize the corresponding log-likelihood function in eq. <3>:

$$L(\lambda) = \log P\left(\mathbf{C}_{u,j-1} \mid S_{u,j}\right) + \log P\left(S_{u,j} \mid S_{u,j-1}\right). \quad <4>$$

We compute the above log likelihood using maximum aposteriori estimation (MAP) to get optimal estimates of the parameters $\lambda_{u,j-1}$. Substituting these optimal weights in eq. <2> along with the actual values of $S_{u,j-1}$ and $\mathbf{C}_{u,j-1}$ gives the conditional probability $P(S_{u,j} | S_{u,j-1}, \mathbf{C}_{u,j-1})$ for the user u at time slice t_j.

4.4 Estimating User Actions

Now we discuss the method of computing the second conditional probability of eq. <1>, that is, the probability of user action given the user state, $p(A_{u,j}|S_{u,j})$.

To estimate this probability, it is sufficient to determine the emission probabilities of actions given the states, that is, $\mathbf{b}_u(A|S)$ learnt for user u over the training time period of T time slice prior to t_0. Since the user action distribution during the training period is known and the states are hidden, we can use a generative model [7] to decode the state sequence given the observed actions and thereby obtain optimal estimates of the emission probabilities $\boldsymbol{b}_u$. Note, using a straight-forward Hidden Markov Model in this case to determine the emission probabilities of actions is not suitable because the state transition probabilities are affected by the user context. Finally we compute the conditional probability of user action given the user state, $p(A_{u,j}|S_{u,j})$ as follows:

$$p\left(A_{u,j} \mid S_{u,j}\right) = \boldsymbol{b}_u\left(A \mid S\right). \quad <5>$$

To conclude this section, we substitute the probabilities from eq. <2> and <5> into eq. <1> to determine the probability of actions for each user u in the social network at each time slice t_j, where $1 \le j \le M$.

5. THE EVOLUTION FRAMEWORK

Synchrony is a temporal phenomenon that persists over a certain time period. Hence online learning methods (e.g. incremental SVM Regression) that incrementally train and predict a value at each time slice, are not helpful in our case – we need estimates of the probability of actions of users in a social network for a set of M future time slices altogether.

To tackle this problem, we propose an evolution framework. First, we need to estimate the size of the social network for each of the M future time slices. This is because it is likely that certain users would leave and some would join over time in performing the specific action. Second, we need to determine which user models to use at the M time slices. Finally we need to determine a measure of the probability of actions for each of the evolved networks to predict an emergent synchrony.

5.1 Estimating the Network Size

Synchrony is likely to involve sustained participation from old users, given by the parameter, sustainability α, as well as attract newer users to perform the action, given by the parameter called the rate of attraction β (section 2). However note that the values of α and β are not available to us over the set of M time slices wherein we intend to predict synchrony. Hence we have to learn the values of α and β in order to estimate the size of the social network at each future time slice. We make a simple conjecture that α and β in a small time window in the past (say, q time slices before t_1, where t_1 is the time slice of the start of synchrony prediction) would reflect how the size of the user set is changing over time. Using the mean rates of α and β over the prior q time slices, we predict the number of users at t_j ($1 \le j \le M$) to be approximately the sum of the number of continuing old users and the new users:

$$\left|U_j\right| \approx \alpha \cdot \left|U_{j-1}\right| + \beta \cdot \left|\bigcup_{i=0}^{j-1} U_i\right|. \quad <6>$$

Given the sizes of the evolved networks over the M future time slices, we discuss the evolution of user models in the next sub-section.

5.2 Evolving User Models

Now we need to determine the specific set of users $\{U_1, U_2, \ldots, U_M\}$ corresponding to each of the evolved networks. Note, the size of network at a time slice t_j could be related to that at t_{j-1} by two relationships – $|U_j| \le |U_{j-1}|$, or $|U_j| > |U_{j-1}|$. We construct the user set for these cases in the following manner:

1. In the first case, U_j comprises the subset of users from U_{j-1} who had the maximum mean probability of actions at t_{j-1}.
2. In the second case, we add those users to U_j who have the maximum probability of actions over all *other* topics at t_{j-1}.

Given the user sets over the M time slices, we now predict their probability of actions at each time slice – that is, evolve the user models. Note our learning framework discussed in section 4 gives estimates of the probability of actions of a user at t_j, given the previous actions and user context $\mathbf{C}_{u,j-1}$ at t_{j-1}. In our set of evolved networks, since the actual values of the contextual attributes are not available for time slices after t_0, we project estimates of the user context over each of the M time slices in order to predict a user's probability of actions.

From section 4 we know that the contextual attributes for the socially aware state is given by, $\mathbf{C}^1_{u,j} = (\eta_{u,j-1}, \omega_{u,j-1}, p(A_{u,j-1}), p(\kappa_{u,j-1}), p(R_{u,j-1}))$ at a certain time slice t_j. The attributes $\eta_{u,j-1}$ and $\omega_{u,j-1}$ can be updated based on the predicted actions of the user u's contacts and those of the seed users in the previous time slice. Similarly $p(A_{u,j-1})$ can be updated based on the predicted value of u's actions. The estimates of the frequency of comments and that of replies from u are not available to us; hence the two attributes $p(\kappa_{u,j-1})$ and $p(R_{u,j-1})$ are held constant over the period of the M time slices. Note, for the contextual attributes $\mathbf{C}^0_{u,j}$ associated with the socially unaware state, we are only able to update $p(A_{u,j-1})$.

Based on the predicted actions and the predicted estimates of the user context, we use our learning framework in section 4 to compute measures of $p(A_{u,j})$ at each time slice t_j, $1 \le j \le M$ and thereby evolve the user models over the M time slices.

5.3 Predicting Synchrony

Based on the evolved networks sizes and evolved user models that comprise these networks, we now use the predicted probability of actions of the user to quantitatively predict an emergent synchrony in the set of M future time slices. Our main idea is that if the mean probability of actions of all users in $\{U_1, U_2, \ldots, U_M\}$ over each of the M future time slices is very high, it implies (quantitatively) an emergent social synchrony in the evolved social networks. In particular, on Digg this implies a large digging probability.

However, note that social synchrony being a social network phenomenon, we are only interested in those users who are engaged in the actions given the socially aware state. Hence the quantitative measure of social synchrony for a certain topic over a set of M future time slices is given by the vector $(p_1, p_2, \ldots, p_M)$ where p_j is the mean probability of actions for all users $u \in U_j$ $(1 \leq j \leq M)$, given u is in the socially aware state.

Now we briefly summarize our computational framework. We proposed a two-stage method (learning and evolution) to predict emergent synchrony in social networks. The learning framework yields estimates of user actions given her context and previous actions. In the evolution framework, we have developed a framework which can evolve network sizes over the set of M future time slices as well as evolve user models to predict the probability of actions of users. These predicted probabilities are then used to predict synchrony.

6. EXPERIMENTAL RESULTS

In this section we present our experimental results comprising a description of the dataset, quantitative analysis of synchrony, comparative study of our method and finally some open issues.

Dataset. The dataset used for the experiments is crawled from the social media Digg. We seeded our crawling from the stories in the featured category 'Popular' on the Digg website. We crawled all stories in this list and which submitted over August and September 2008. We identified the unique users from these set of stories and constructed their degree distribution (i.e. the number of contacts). From the degree distribution, we picked a set of 500 users with the highest degrees. We crawled all the stories, diggs, comments, replies submitted by them over the two months and collected their contacts. We used a snowballing technique, i.e. iteratively followed this procedure for a set of 21,919 users. In total, this dataset comprises 187,277 stories, 7,622,678 diggs, 687,616 comments and 477,320 replies over a set of 51 topics in this time range.

6.1 Quantitative Analysis

Now we present a quantitative analysis of our prediction results from two different perspectives – prediction of synchrony and relationship of certain properties of the seed set to synchrony.

6.1.1 Prediction of periods of synchrony

We discuss an analysis of the effectiveness of our model in predicting user actions over time, and relationship to presence / absence of synchrony over a set of topics. We focused on six different topics – 'US Elections', 'World News', 'Olympics', 'Comedy', 'Celebrity' and 'Tennis' to demonstrate the performance our prediction. Each of these topics has a total number of 14,245, 23,935, 10,732, 8,356, 4,735 and 6,774 users respectively. The rationale behind choosing these topics was to test our model on a diverse range of topics – some of which were current at that point in time (e.g. 'US Elections'), as well as some of which were consistently discussed topics (e.g. 'Comedy').

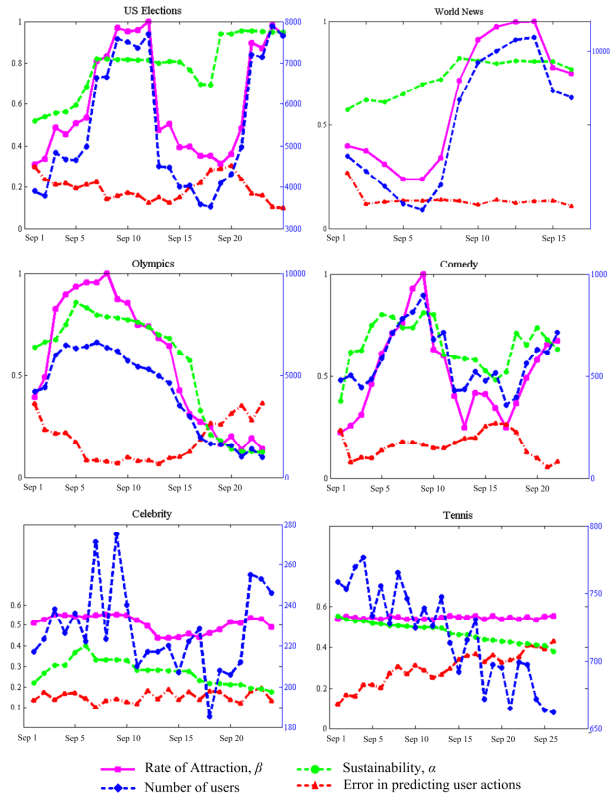

Figure 7: Prediction of user actions based on our framework for six topics: 'US Elections', 'World News', 'Olympics', 'Comedy', 'Celebrity' and 'Tennis'. Three variables – our predicted error, α and β (normalized) are shown on the left Y-axis, while the number of users is shown on the right Y-axis. Note, our model predicts the user actions with low mean error rate.

Experiments were conducted over the time period of the testing data taken to be September 2008, with training data over the month of August 2008. At each day in the test set, we predicted the user actions for future M days based on a time slice duration of t_j days. Note that choosing reasonable values of t_j and M are extremely important because continual evolution of network and user models over several time slices (without retraining) could soon yield very high prediction error rates of the probabilities of user actions. We therefore chose an empirical threshold of prediction error equal to 20%, and our goal was to determine a value pair t_j, M such that the error was no more than 20%. Obviously we want M to be as high as possible, and error to be still less than 20%. To determine such a pair, for each topic, we

considered different durations of time slices t_j= 0.5, 1, 2, 3 and 4 days and predicted the user actions for different values of M= 1-20 days. The optimal durations of time slices were found to be 2, 3, 2, 3, 1 and 2 days respectively for each of the six topics. And the corresponding optimal values of M were 6, 18, 8, 9, 7 and 5 days respectively.

The results of prediction based on these values are shown in Figure 7. Four different variables are shown in each plot in the figure – the actual rate of attraction β, sustainability α, the actual number of users and the error on predicting user actions. The first four topics indicate the presence of synchrony in certain time regions. This can be understood from the following observations – the number of users is high in certain time periods than others and the rate of attraction β and sustainability α are also reasonably large (twice as large as the previous value). For example, synchrony is observed from Sept 6 to Sept 13 for the topic 'US Elections', while from Sept 8 to Sept 11 for the topic 'World News'.

We now observe the performance of our framework in predicting user actions during these periods of social synchrony. We notice that our model performs very well in prediction after the onset of synchrony (≤15% error), compared to the beginning and the end (~20% error). This can be accounted for by the fact that our model learns the values of α and β based on a short time span in the past. At the onset of the synchrony, the total number of users and the values α and β exhibit rapid increase. Hence their estimates based on the past do not match well with the actual values, affecting our prediction performance. Similarly towards the end of synchrony our model is trained on the rapidly rising values of α and β and it is not it is not able to account for the decrease in the two values. In the case of topics without synchrony, e.g. 'Celebrity' and 'Tennis', our framework is able to predict user actions with sufficiently low error rates.

To summarize, we observe that our model on average is consistently able to predict user actions with low error (15.2±4.3%) during time periods with and without synchrony over all the 51 topics present in the crawled dataset.

6.1.2 Properties of seed users

We now investigate if network properties of the seed user set, including size, degree distribution, clustering coefficient and measure of vulnerability (i.e. the weighted sum of the degrees of all individual users in the seed set) are correlated with the phenomenon of synchrony.

Table 1 shows the measures of the different properties of the seed set for six topics. For the first four topics where synchrony is present, the mean of the degree distribution is high, variance is low, clustering coefficient is low and the vulnerability measure is high. This is due to the fact that during synchrony, the network size evolves very rapidly and hence the seed set of users does not form a cohesive sub-network – hence the low clustering coefficient. High measure of vulnerability indicates high degrees of users, which further indicates that the action of digging is observable to a large set of users (through feeds over their contacts), triggering several new users to mimic the action, leading to an emergent synchrony. However the topics 'Celebrity' and 'Tennis' do not exhibit synchrony. In this research hence we have preliminary evidence that the properties of the seed set – high mean and low variance of degree distribution and low clustering coefficient are likely to be correlated with an emergent synchrony at a later point in time.

Table 1: Properties of the seed user set for six different topics. $|U_0|$ is the size of seed set, μ_0, σ_0 are the mean and variance of the degree distribution of U_0, ρ_0 is the clustering coefficient of the seed users' sub-network and ς_0 is the measure of vulnerability.

Topics	$\|U_0\|$	μ_0	σ_0	ρ_0	ς_0	Synchrony
US Elections	**334**	**42**	**0.13**	**0.21**	**0.72**	*Yes*
World News	**582**	**37**	**0.18**	**0.11**	**0.76**	*Yes*
Olympics	**121**	**28**	**0.21**	**0.23**	**0.67**	*Yes*
Comedy	**79**	**27**	**0.17**	**0.26**	**0.60**	*Yes*
Celebrity	42	11	0.61	0.53	0.23	*No*
Tennis	64	14	0.44	0.54	0.13	*No*

6.2 Comparative Study

We now compare the performance of our model, i.e. the error predicting user actions (section 6.1.1) based on four baseline methods. The techniques consist of two non-social network based methods and two social network based methods. The first is a simple temporal trend learning method of user actions. The second is a linear regressor based method which uses the coefficients associating a user's activities like commenting and replying with her actions to predict actions in the future. The third baseline is the SIR (susceptible-infected-removed) epidemiological model [4] popularly used in determining information cascades in social networks. The last baseline is a threshold based model of global cascades [11] based on the idea that a user participates in an action by changing her state to 'active' only when a certain sufficiently large fraction of her contacts have already done so. Note that these baseline methods are used in an online learning setting to train and predict the probabilities of the actions over the future. We believe that this constructs a more rigorous comparative benchmark instead of extrapolating the probabilities over the set of future time slices.

Table 2: Error in predicting user actions based on our method against baseline techniques – trend learning method (B_1), user activity regression based framework (B_2), SIR based epidemiological model (B_3) and a simple threshold based model of global cascades (B_4). Our method performs the best among all baseline techniques.

Topics	Our Method	B_1	B_2	B_3	B_4
US Elections	**0.19**	0.67	0.52	0.38	0.35
World News	**0.11**	0.41	0.36	0.29	0.28
Olympics	**0.19**	0.54	0.49	0.44	0.41
Comedy	**0.13**	0.46	0.4	0.31	0.27
Celebrity	**0.12**	0.49	0.36	0.29	0.22
Tennis	**0.15**	0.53	0.41	0.32	0.27

The results of evaluation against the four baseline methods have been shown in Table 2. The errors in prediction are shown for the same six different topics and over the time period of the testing data – September 2008. The evaluation results of our method against baseline techniques reveals that it performs the best among all with a mean improvement in prediction ranging 18-56% over all the 51 topics in the dataset. The trend based learning method (B_1) and the regressor based user activity measure (B_2), both perform worse than the other two social network based baseline methods. This explains that synchrony is a social network phenomenon and simple time series based prediction methods naturally yield poor results. The performances of the SIR epistemological model (B_3) and the global cascades model (B_4) are similar; however the cascades model yields slightly lower error. This indicates that the local network topology, that is, the contacts of a user affect her decision to perform an action. Our method has two key improvements over the baseline techniques – (a) the ability to predict user actions for a set of time slices into the future unlike online learning, and (b) ability to take into account the dynamically changing sizes of the user set via the rate of attraction and sustainability. They help boosour prediction performance.

6.3 Open Issues

Among the open issues in our approach, note that the notion of synchrony is based on prediction of user actions over a set of time slices in the future. However, determining the exact size of the time horizon (i.e. the number of time slices into the future we can predict) is challenging. This is because typical periods of synchrony could differ for different topics as well as could depend upon contextual factors which are directly not observable in the social network. Moreover the socially aware and unaware states of a user could be affected by her intrinsic preferences – for example, Alice diggs political news because her friends blog about them in the Blogosphere. These factors are not captured by our representation of user context.

7. CONCLUSIONS

In this paper, we developed a computational framework to predict emergent social synchrony in social media. Synchrony was defined to have the following characteristics – a given topic and an agreed upon user action, notion of a *seed* user set, and the rate of attracting new users as well as sustained participation of earlier users. Our method consisted of two parts – the learning framework and the evolution framework. In the learning framework, we developed a DBN based representation to predict the probability of user actions over a set of time slices into the future. In the evolution framework, we evolved the social network size and the user models over a set of future time slices to predict social synchrony. We conducted extensive experiments on a large dataset crawled from the popular news sharing social media site Digg. The results showed that our model yielded low error (15.2±4.3%) in predicting user actions (or diggs) during time periods with and without synchrony. Comparison with different baseline techniques indicated that our method showed significant improvement in predicting user actions.

For future work, we intend to develop sophisticated theoretical models that can identify the conditions triggering social synchrony. The role of social synchrony in citation networks can be useful to identify time periods of paradigm shifts as well. We are also interested to explore how the action of *burying* a news article on Digg (apart from digging) affects the emergence of a synchrony in the future.

8. REFERENCES

[1] *Digg* http://digg.com/.

[2] E. ADAR and L. A. ADAMIC (2005). *Tracking Information Epidemics in Blogspace*. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society**:** 207-214.

[3] A. ANAGNOSTOPOULOS, R. KUMAR and M. MAHDIAN (2008). *Influence and correlation in social networks*. Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. Las Vegas, Nevada, USA, ACM**:** 7-15.

[4] M. CHA, A. MISLOVE, B. ADAMS, et al. (2008). *Characterizing social cascades in flickr*. Proceedings of the first workshop on Online social networks. Seattle, WA, USA, ACM**:** 13-18.

[5] Y. FERNANDESS and D. MALKHI (2008). *On spreading recommendations via social gossip*. Proceedings of the twentieth annual symposium on Parallelism in algorithms and architectures. Munich, Germany, ACM**:** 91-97.

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

[7] K. C. JUNG, S. M. YOON and H. J. KIM (2000). *Continuous HMM applied to quantization of on-line Korean character spaces.* Pattern Recogn. Lett. **21**(4): 303-310.

[8] J. LESKOVEC, L. A. ADAMIC and B. A. HUBERMAN (2006). *The dynamics of viral marketing*. Proceedings of the 7th ACM conference on Electronic commerce. Ann Arbor, Michigan, USA, ACM**:** 228-237.

[9] J. LESKOVEC, A. KRAUSE, C. GUESTRIN, et al. (2007). *Cost-effective outbreak detection in networks*. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. San Jose, California, USA, ACM**:** 420-429.

[10] J. LESKOVEC, M. MCGLOHON, C. FALOUTSOS, et al. (2007). *Cascading behavior in large blog graphs: Patterns and a model*, Society of Applied and Industrial Mathematics: Data Mining,

[11] D. WATTS (2002). *A simple model of global cascades on random networks.* Proceedings of the National Academy of Sciences **99**(9): 5766-5771.

[12] C. ZHAI and J. LAFFERTY (2004). *A study of smoothing methods for language models applied to information retrieval.* ACM Trans. Inf. Syst. **22**(2): 179-214.