Modeling Information Diffusion in Implicit Networks

Jaewon Yang
EE Department, Stanford University
crucis@stanford.edu

Jure Leskovec
CS Department, Stanford University
jure@cs.stanford.edu

Abstract—Social media forms a central domain for the production and dissemination of real-time information. Even though such flows of information have traditionally been thought of as diffusion processes over social networks, the underlying phenomena are the result of a complex web of interactions among numerous participants.

Here we develop a *Linear Influence Model* where rather than requiring the knowledge of the social network and then modeling the diffusion by predicting which node will influence which other nodes in the network, we focus on modeling the global influence of a node on the rate of diffusion through the (implicit) network. We model the number of newly infected nodes as a function of which other nodes got infected in the past. For each node we estimate an influence function that quantifies how many subsequent infections can be attributed to the influence of that node over time. A nonparametric formulation of the model leads to a simple least squares problem that can be solved on large datasets.

We validate our model on a set of 500 million tweets and a set of 170 million news articles and blog posts. We show that the Linear Influence Model accurately models influences of nodes and reliably predicts the temporal dynamics of information diffusion. We find that patterns of influence of individual participants differ significantly depending on the type of the node and the topic of the information.

I. Introduction

The information we experience comes to us continuously over time, assembled from many small pieces, and conveyed through social networks as well as other means. The merging of information, network structure, and flow over time opens interesting questions about the large-scale behavior in information networks.

Even though the diffusion of information has been an active research area recently [7], [12], [26], [28], modeling the diffusion in social networks has proven to be a challenging task. It is difficult to obtain large scale diffusion data and to identify and track on a large scale the elements, such as recommendations [25], links [27], [28], tags [8], [7], topics [3], phrases or "memes" [26], that spread and propagate through networks. Even if one does obtain large scale real-world diffusion data, however, the issue of modeling the underlying process still remains. Traditionally, models of diffusion and cascading behavior have formalized the spread of ideas, information and influence as processes taking place on social and information networks [13], [15], [31], where each individual node is either *active* (infected, influenced) or *inactive*, and active nodes can then spread the *contagion* (information, influence, disease) along the edges of the underlying network. Parameter estimation of such models is challenging due to the heterogeneity of the nodes and data sparsity. Only recently has the availability of large social network and corresponding diffusion data made it possible to estimate such models in practice [14], [30].

When using such models and fitting them to real-world data one makes several assumptions: (a) complete network data is available, (b) contagion can only spread over the edges of the underlying network, (c) the structure of the network itself is sufficient to explain the observed behavior. However, in many scenarios, the network over which diffusion takes place is in fact implicit or even unknown. Commonly, we only observe when nodes got "infected" but not *who* infected them. In case of information propagation, people usually discover new information without explicitly acknowledging the source. In word of mouth and viral marketing settings, we only observe people purchasing products or adopting new behaviors without explicitly knowing who was the influencer that caused the adoption or the purchase. Similarly, in virus propagation, we observe people getting infected without knowing who infected them. Moreover, many times an activation of a node is not just a function of the social network but also depends on many other factors like imitation and recency. For example, people prefer the most recent information, and they discover new information or make decisions by using many different means, like the search engines, media sites, online forums and blogs or employing their social networks. Thus, even though flows of information and influence have traditionally been thought of as diffusion processes over underlying social networks [13], [15], [29], [31] existing models and formulations may be too constrained to capture the complexity of the underlying phenomena.

Modeling diffusion and temporal variation. Here we address the above issues by developing a model of diffusion where no explicit knowledge of the network is necessary. Rather than predicting which node in the network will infect which other nodes, we focus on modeling the global influence a node has on the rate of diffusion through the (implicit) network. Models of diffusion generally ignore time and operate in discrete epochs. Instead, we accurately model

not only the influence each node has on the diffusion but also how the diffusion unfolds over time.

Consider the diffusion of information in online media, where no explicit network of who spreads the information to whom exists. As the information propagates, a blogger or a website gets "infected" when it mentions the information. In such cases individuals and websites may act in diverse ways: News wire services play an amplifying role, blogs can serve both as early detectors and elaborators (or echo chambers), while the mainstream media imparts a dominant force in the direction the news cycle takes [23], [16]. For example, some websites may act as "influentials" or early adopters [32]. Bloggers and mainstream media are pushing new content into the system in different manners [22], [11], and often the content generated by blogs is regarded to be more credible than that from the mainstream media [21].

In this paper we aim to develop an understanding of the mechanisms by which the rate of diffusion rises and decays over time. What causes certain information cascades to grow large and why others remain small? And, what are the roles of different participants in the dynamics of diffusion?

Linear Influence Model (LIM). We consider the temporal variation in a diffusion-based framework and build on the view adopted by the literature on social influence [10], [20]. We formulate the *Linear Influence Model (LIM)* by starting with the assumption that the number of newly infected nodes depends on which other nodes got infected in the past. We then model the number of newly infected nodes as a function of the times when other nodes got infected in the past. In this model, each node has an *influence function* associated with it. Then the number of newly infected nodes at time t is a function of influences of nodes that got infected before time t. Going back to our example of information diffusion, we assume that the number of websites (i.e., nodes) that mention particular information depends on which other websites mentioned the information beforehand. Then one can view the website's influence function as follows: after website u mentions the information at time t, this causes additional $I_u(1)$ other sites to mention the information in the next time step, $I_u(2)$ new mentions after two time steps, and so on.

We show that node influence functions can be efficiently estimated by formulating a regression task where the goal is to learn an influence function $I_u(t)$ for each node u such that the overall number of newly infected nodes at time t is the sum of influences of previously infected nodes. We model influence functions in a non-parametric way and show that they can be estimated using a simple least squares procedure.

We experiment on two massive real world datasets: a corpus of 500 million Twitter posts, and a set of 172 million news articles. We model the information diffusion in these two datasets by estimating node influence functions. Experiments show that our model outperforms standard time series forecasting methods when predicting the magnitude

and the rate of information diffusion. We find that influence functions exhibit distinct shapes depending on the node type (newspaper, news agency, blog), and the topic of information. We also find that Twitter users who have the most followers are not the most influential in terms of information propagation.

Applications. Estimating the influence of a node on the diffusion process is important as it gives us a direct way to quantify patterns of influence and roles different nodes play in the diffusion of various types of contagions (topics of information, types of products). The model allows us to predict the future adoption of the contagion and to quantify the relative influence of nodes, and thus helps us answer questions such as: What is the influence of a particular node? How does its influence change over time?

Even though we present our model in the context of the information diffusion and adoption in social media, our work is also applicable to many other settings. Most generally, we can think of a contagion (information, virus, innovation) that is spreading through the network but we only observe its volume (the number of newly infected nodes) over time. Now, based on the times when a small number of nodes got infected by the contagion we model the influence of these nodes on the overall volume and the temporal dynamics of the diffusion. This setting naturally applies to viral marketing [6], [18], where we observe people purchasing products or adopting particular behavior without explicitly knowing who was the influencer. Thus, for viral marketing, estimating the influence functions (i.e., how many subsequent purchases a node influences) is of considerable interest. Similarly, in epidemiology and virus propagation, we observe people getting sick without usually knowing how they got infected [4]. Here our model allows us to estimate the number of subsequent infections produced by each node without the knowledge of the network.

II. PROPOSED METHOD

Next we formally introduce the *Linear Influence Model (LIM)*. Even though our model is widely applicable, we restrict our discussion to the setting of information diffusion in online media, where we track nodes (blogs, mainstream media, or users on Twitter) mentioning particular pieces of information (Twitter hashtags, or short textual phrases).

Model formulation. Consider a set of nodes that participate in a diffusion process. As the information diffuses, nodes become "infected" when they adopt (mention) the information. We consider the setting where we observe only the time t_u when a particular node u mentioned the information and do not require the knowledge of the underlying network. We define the *volume*, $V(t)$, as the number of nodes that mention the information at time t. We aim to model the volume over time as a function of which other nodes have mentioned the information beforehand.
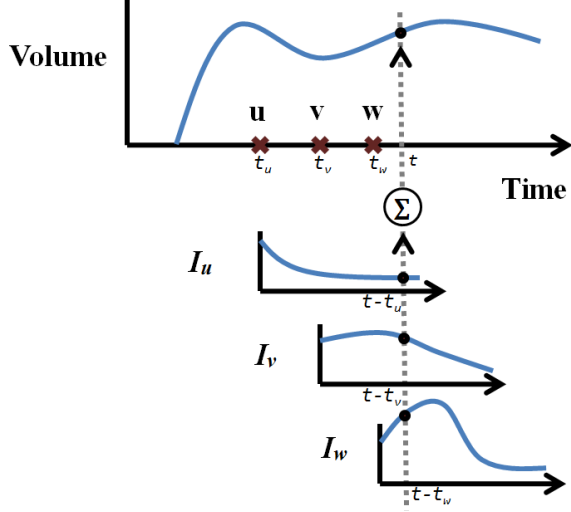
Figure 1. The Linear Influence Model models the volume of diffusion over time as a sum of influences of nodes that got "infected" beforehand.

We posit that each node u has a particular non-negative *influence function* $I_u(l)$ associated with it. One can simply think of $I_u(l)$ as the number of followup mentions l time units after node u adopted the information. Or equivalently, after node u mentions the information, this triggers an additional $I_u(1)$ mentions in the next time step, $I_u(2)$ mentions after two time steps, and so on. Now, we aim to model the relation between the volume $V(t)$, and the influence functions of nodes u that mention the information at times t_u ($t_u < t$). We simply assume that the volume $V(t)$ is the sum of properly aligned influence functions of nodes u:

$$V(t+1) = \sum_{u \in A(t)} I_u(t - t_u)$$

where $A(t)$ denotes the set of already active (infected, influenced) nodes u that got activated prior to time t ($t_u \le t$).

Figure 1 illustrates the model. The curve on the top represents the volume $V(t)$ over time, and t_u, t_v, and t_w denote the times when nodes, u, v and w, got infected. After the nodes got infected, they each influence additional $I_u(t - t_u)$, $I_v(t - t_v)$ and $I_w(t - t_w)$ infections at time t. So the volume $V(t)$ at time t is the sum of the influences of the three nodes.

A natural question then is how to model the individual influence functions $I_u(l)$. There are two general approaches. The first is a parametric approach, where one could assume that functions $I_u(l)$ follow a certain parametric form, such as an exponential $I_u(l) = c_u e^{-\lambda_u l}$ or a power law $I_u(l) = c_u l^{-\alpha_u}$ with parameters depending on the node u. Although such a model would be very clean and simple, it has an important drawback, as it assumes that the influence functions of all nodes follow the same parametric form. This assumption may be too simplistic to capture the complex dy-

namics of diffusion. This is especially true in online media, where a diverse set of participants (blogs, newspapers, TV stations, news agencies) play very different roles and have very different impacts on the overall dynamics of diffusion.

To account for this diversity we use a non-parametric approach. This way we do not make any assumptions about the shape of the influence functions and we let the model estimation procedure find the most appropriate shapes. We achieve this by considering the time to increase in discrete intervals (e.g., one hour). Then we can represent an influence function $I_u(l)$ as a non-negative vector of length L, where l^{th} value represents the value of $I_u(l)$. Setting the length of vector I_u to L simply means that the influence of a node drops to zero after L time units.

Such non-parametric formulation of the Linear Influence Model makes no assumptions about the shape of individual influence functions. This offers great modeling flexibility, as different nodes can have very different patterns of influence. Furthermore, we can study how the shape of the influence functions varies for different types of nodes or for different types of contagions (e.g., textual phrases of different topics). Finally, nodes can be grouped based on the shape of their influence functions to gain further insights into the roles different nodes play in the diffusion process.

Model parameter estimation. Next we present an efficient procedure to estimate parameters (i.e., influence functions) of the LIM model. Consider a set of N nodes and the data on how K different contagions diffused between the nodes over time, where each contagion can infect any arbitrary subset of nodes. We then represent this data as a large indicator function $M_{u,k}(t)$, where $M_{u,k}(t) = 1$ if node u got infected by contagion k at time t, and 0 otherwise. Note that the volume $V_k(t)$ of contagion k at time t is simply defined as the number of nodes that got infected by k at time t. We then model the volume $V_k(t)$ as a sum of influences of nodes u that got infected *before* time t:

$$V_k(t+1) = \sum_{u=1}^{u=N} \sum_{l=0}^{l=L-1} M_{u,k}(t-l) I_u(l+1) \quad (1)$$

The first summation goes over all the nodes, while the second goes over the time-length of influence functions. Given the current time t we first check whether node u got infected with contagion k l-time units ago. If so, then $M_{u,k}(t-l) = 1$ and node u contributes its influence of $I_u(l)$ to the total volume.

Next, we show how to estimate the influence functions $I_u(t)$ that most accurately predict volume $V_k(t+1)$ given the particular other nodes that got infected in the past. Generally, we will not be interested in estimating the influence functions of all the nodes but will rather model the total volume $V(t)$ as a function of a small set of N nodes of interest. Thus, $V(t)$ models the total volume over the whole universe of nodes (all online media, all Twitter users, etc.),
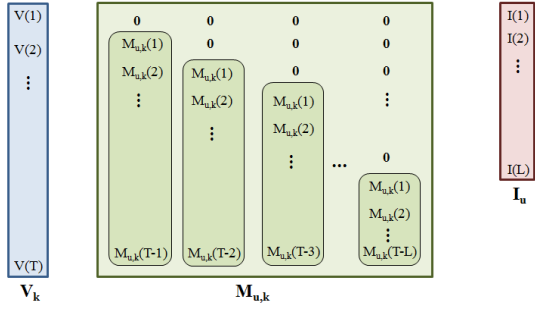
(a) Volume vector $\mathbf{V}_k$ of length T, influence vector $\mathbf{I}_u$ of length L, and a $T \times L$ lower-triangular block $\mathbf{M}_{u,k}$ of influence indicator matrix $\mathbf{M}$
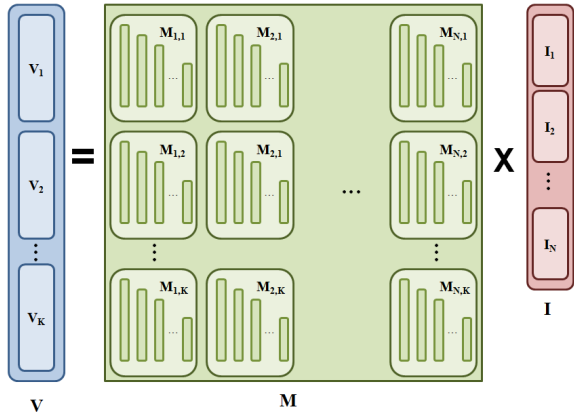
(b) Vector $\mathbf{V}$ of length $K \cdot T$, vector $\mathbf{I}$ of length $L \cdot N$, and an influence indicator matrix $\mathbf{M}$ of size $K \cdot T \times N \cdot L$.

Figure 2. The structure of the matrix equation $\mathbf{V} = \mathbf{M} \cdot \mathbf{I}$

while N denotes a small subset of nodes of interest (e.g., only newspapers, or a small subset of most active Twitter users). We also assume that the number of contagions is larger than the number of nodes of interest ($K > N$).

Since time t increases in discrete intervals, we represent $V_k(t)$, $M_{u,k}(t)$, and $I_u(l)$ as vectors and matrices, and formulate a least squares-like problem, where for each node of interest u the goal is to estimate L values of its influence function, $I_u(1), \ldots, I_u(L)$. We show that values $I_u(l)$ can be estimated by a simple matrix equation using the fact that volume $V_k(t)$ is a linear function of influence $I_u(l)$ (Eq. 1).

To formulate the matrix equation we first define the *volume vector* $\mathbf{V}$, the *influence vector* $\mathbf{I}$, and the *influence indicator matrix* $\mathbf{M}$ (Figure 2). We compose a column vector $\mathbf{V}$ of length $K \cdot T$ by simply thinking of volume $V_k(t)$ of each contagion k as a vector V_k of length T indexed by t, and then concatenating the contagions for $k = 1, \ldots, K$. Second, we compose an influence vector $\mathbf{I}$ of length $N \cdot L$ by considering each $I_u(l)$ as a vector I_u of length L indexed by l, and then concatenating them. Last, we compose a binary influence indicator matrix $\mathbf{M}$ of $K \cdot T$ rows and $N \cdot L$ columns. Consider that node u, got infected with contagion k at time t. Then we set entries (i, j) of matrix $\mathbf{M}$ to 1 for $i = kT(t+l)$

and $j = uL(t+l+1)$, where $l = 0, \ldots, \min(L-1, T-t)$. Note that matrix $\mathbf{M}$ has a block structure where every block $\mathbf{M}_{u,k}$ represents a node–contagion pair, and if a node u got infected by contagion k at time t this creates a diagonal stripe of ones in a block $\mathbf{M}_{u,k}$, i.e., $\mathbf{M}_{u,k}[t+l, l+1] = 1$, for $l = 0, \ldots, L-1$ (Fig. 2(a)). This way the T rows of $\mathbf{M}_{u,k}$ account for time and the L columns for how the influence of a node changes (up to L time units) after it got infected.

Now our aim is to solve a matrix equation $\mathbf{V} = \mathbf{M} \cdot \mathbf{I}$ where we aim to estimate values of the influence vector $\mathbf{I}$ given the values of the volume vector $\mathbf{V}$ and the influence indicator matrix $\mathbf{M}$. However, due to noise and the fact that the system is over-determined ($K \cdot T \gg N \cdot L$) we do not expect that an exact solution exists. Thus we aim to find the $\mathbf{I}$ that minimizes the prediction error measured by the Euclidean distance between the true and the predicted volume:

$$\text{minimize} \quad ||\mathbf{V} - \mathbf{M} \cdot \mathbf{I}||_2^2$$
$$\text{subject to} \quad \mathbf{I} \geq 0$$

where $|| \cdot ||_2^2$ denotes the squared Euclidean norm.

The above optimization problem is called a non-negative least squares (NNLS) problem [24] and can be solved efficiently even for a large number of nodes and contagions. The sparse nature of the influence indicator matrix $\mathbf{M}$ helps to further expedite the calculation. We use the Reflective Newton Method [9] which takes less than a second to solve a problem with $K = 1{,}000$, $L = 10$, $T = 120$, and $N = 100$. In practice we also apply the Tikhonov regularization [19], which has the effect of smoothing the non-parametric estimates.

Extensions: Accounting for novelty. So far, we have assumed that a node has the same influence regardless of how early or late in the diffusion they appear. This means that the influence of a node is same even if it mentions the information very early or very late. However, nodes are more likely to adopt novel and recent information while ignoring old and obsolete information. In order to account for this effect of recency and novelty [33] we introduce a multiplicative factor $\alpha(t)$ that models how much more/less influential a node is at the time when it mentions the information. We refer to this model as α-LIM:

$$V_k(t+1) = \alpha(t) \sum_{u=1}^{u=N} \sum_{l=0}^{l=L-1} M_{u,k}(t-l) I_u(l+1)$$

Note that $\alpha(t)$ is the same over all contagions. We expect $\alpha(t)$ to start low, quickly peak and then slowly decay. The influence of nodes just before the peak attention will be boosted simply because the information is new and nobody knows about it yet. As time goes by the novelty decays and the benefit of appearing early in the diffusion wears off ($\alpha(t)$ decreases).

In order to estimate $I_u(l)$ and the T values of vector $\alpha(t)$ we observe that the resulting matrix equation is convex

both in $I_u(l)$ when $\alpha(t)$ is fixed and in $\alpha(t)$, when $I_u(l)$ is fixed. Thus we use a coordinate descent procedure, where we iterate between fixing $\alpha(t)$ and solving for $I_u(l)$, then fixing $I_u(l)$ and solving for $\alpha(t)$.

Extensions: Accounting for imitation. Another aspect of information diffusion and adoption is the effect of imitation [26], where nodes imitate one another because the information is popular and everyone talks about it. We refer to the contribution of the imitation as the *latent volume* in a sense that this volume is caused not by influence, but by other factors. We model the latent volume with an additive factor $b(t)$ and refer to the model as the B-LIM model:

$$V_k(t+1) = b(t) + \sum_{u=1}^{u=N} \sum_{l=0}^{l=L-1} M_{u,k}(t-l)I_u(l+1)$$

B-LIM is linear in $I_u(l)$ and $b(t)$, and thus we can use a matrix formulation similar to the one in Figure 2.

Discussions and further extensions. Another direction for extensions is to introduce an additional parameter for each contagion to explicitly model for the attractiveness of different contagions, arguing that some contagions are a priori more interesting, attractive and easier to diffuse. An alternative approach would be for nodes to have multiple influence functions depending on the type or topic of the contagion.

Last, we note that our model can be used for "prediction" and as well as "explanation." So far we have introduced the model in the prediction setting, where we observe a small subset of N nodes that got infected up to time t and want to predict the total volume over *all* nodes in the future time $t+1$. However, we can also use the model for explanation in the sense that we observe a small number of nodes N that got infected up to time t, and we are then interested in predicting the total volume at the current time t. This formulation does not predict (forecast) the total volume in next time step but it rather predicts the total volume at the current time step based on which nodes are currently infected. We consider both formulations to be interesting and valid; however, in the rest of the paper we only consider the predictive formulation, where we aim to predict the future total volume at time $t+1$ based on observing which nodes got infected in the past.

III. EXPERIMENTS

In this section, we evaluate the performance of LIM on two different datasets. We first describe the datasets and the experimental setup, and then evaluate LIM on a time series prediction problem.

Dataset description. First, we consider modeling the diffusion of short textual phrases over the online media space. We apply the Memetracker [26] methodology and extract 343 million short textual phrases from a set of 172 million news articles and blog posts collected from more than 1

million online sources between September 1 2008 to August 31 2009. To ensure that we observe the complete lifetime of a phrase, we only keep phrases that first appeared after September 5. We choose 1,000 phrases with highest volume in a 5 day window around their peak volume. For each phrase, we track which websites mention it during 5 days around its peak volume.

Second, we analyze the diffusion of hashtags on Twitter. Twitter users often tag posts with "hashtags" (e.g., *#iknowsomeonethat*, *#ilovelifebecause*). The emergence and adoption of hashtags create global cascades in the Twitter network. We collect a stream of 580 million Twitter posts (40-50% of all posts) between June 2009 and February 2010. We identify 6 million different hashtags, and then discard hashtags that do not experience a significant peak in their volume (e.g., ♯*musicmonday* and ♯*goodmorning*). We then select 1,000 highest total volume hashtags during the 5 days around their peak volume. As Twitter users adopt at most 1% of the hashtags, we mitigate this data sparsity issue by grouping users into groups of 100 users. We consider 100 groups and model each group as a node. We then model the collective behavior of each group by aggregating all the mentions within the group.

Experimental setup. Volume $V_k(t)$ of a contagion k can naturally be viewed as a time series. We thus evaluate our LIM model on a time series prediction task, where we observe the nodes that got infected with k up to time t and aim to predict the volume $V_k(t+1)$ of the contagion k at future time $t+1$.

To evaluate the model we employ 10-fold cross validation. We split contagions (hashtags, memes) into 10 folds, use 9 folds to estimate the model parameters and evaluate on the remaining fold. For each contagion k in the evaluation fold, we predict the volume $\hat{V}_k(t+1)$ of contagion k at time $t+1$. We then measure the difference between the true volume and predicted volume, $E_k(t+1) = V_k(t+1) - \hat{V}_k(t+1)$, and report the relative error $\sqrt{\sum_{k,t} E_k(t)^2}/\sqrt{\sum_{k,t} V_k(t)^2}$.

In all of our experiments, we use $K = 1,000$ contagions, one hour as the time unit, and set $L = 10$ (i.e., influence of a node decays to zero after 10 hours). Since contagions have very short life spans, we set the length of the volume time series to 5 days (i.e, $T = 120$). For each contagion, we set the start ($t = 0$) of $V_k(t)$ to be first time when the volume of the contagion is twice the average volume in previous 5 time-steps. This has the effect that we start to observe the volume of the time series just before it starts to peak (see Figure 5(a) for example). We also allow each node to mention the phrase or hashtag multiple times during a time unit, i.e., $M_{u,k}(t)$ can be more than 1.

We model the total volume $V_k(t)$ of a contagion based on the influence of $N = 100$ nodes. Note that this is an extremely small fraction of the total number of nodes. In Memetracker, for example, we have more than 1 million
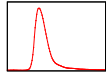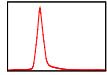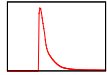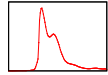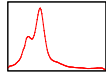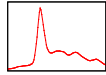
Model							ALL
AR	6.82%	7.08%	8.43%	7.21%	8.47%	8.30%	7.41%
ARMA	6.65%	7.71%	8.29%	6.85%	8.07%	8.71%	7.75%
LIM	13.89%	12.42%	11.41%	20.06%	6.22%	6.24%	14.31%
B-LIM	15.38%	15.19%	12.24%	21.27%	8.15%	6.99%	15.71%
α-LIM	15.50%	14.59%	11.50%	20.08%	7.13%	6.71%	15.26%

Table I

REDUCTION IN PREDICTION ERROR OVER 1-TIME LAG PREDICTOR ON MEMETRACKER DATA FOR SIX TYPES OF SHAPES OF VOLUME OVER TIME.

different websites (nodes) that participate in the diffusion and we aim to model the total number of sites that will mention the phrase based on the information about mentions from only 100 highest volume sites. Similarly, in Twitter we model the hashtag volume over the 25 million active users based only on the information about 10,000 users, which is only 0.04% of the total active users.

Baseline methods. We compare the performance of LIM with three time series prediction methods. First, a 1-time lag predictor simply takes the volume at the current time as the prediction for the volume at the next time, $\hat{V}_k(t+1) = V_k(t)$. We also consider two standard time series regression methods: the Autoregressive Model (AR), and the Autoregressive Moving Average Model (ARMA) [5] both of order L. The AR model is equivalent to a special case of LIM where we assume that all the nodes have the same influence function. ARMA uses AR with an additional ingredient, the moving average model. We use training folds to estimate model parameters, and then evaluate on the test fold, where we predict the volume at time $t + 1$ given the time series of volume $V_k(t)$ up to time t.

Time series prediction problem. We evaluate our LIM model on the task of predicting the volume of a contagion over time. We evaluate three versions of the LIM model (i.e., LIM, B-LIM, and α-LIM) and compare the performance with the three time series forecasting methods (1-time lag predictor, AR and ARMA). The purpose of these experiments is not to build a perfect time series predictor. Rather, we aim to evaluate whether the modeling assumptions of LIM are reasonable and to what degree the observed dynamics of diffusion can be attributed to the influence of nodes.

Table I shows the relative reduction in error over the 1-time lag predictor on the Memetracker data for all phrases, and also for phrases grouped based on the shape of the volume over time [2]. While AR and ARMA give 7.5% improvement, LIM and its variants outperform AR and ARMA by a factor of two. We find the results to be similar for predicting the adoption of Twitter hashtags (table not shown for brevity) where AR and ARMA give about 1% improvement over 1-time lag predictor, while LIM gives 6.1% error reduction (B-LIM 6.3%, α-LIM 3.5%).

There are several interesting observations about these results. First, notice that AR is equivalent to LIM with the same influence function for all nodes. Our results suggest that nodes have very different levels of influence and that we obtain a substantial benefit from the non-parametric approach. Moreover, we also observe that LIM gives better results for modeling the adoption of textual phrases in online media than for modeling the adoption of Twitter hashtags. These results suggest that there are a relatively small number of media sites that have large influence on the adoption of textual phrases, while the influence of top Twitter users on the adoption of Twitter hashtags is smaller. These results align well with the two-step theory of information flow [22], which has been developed in sociology to reconcile the role of the media with the observation that in many scenarios individuals are influenced by the neighbors in their social networks as well as by the mainstream media. The theory is called a "two-step flow" as the information and influence "flows" from the mass media through opinion leaders to the public. In our context here, the results suggest that while the media space is occupied by relatively few very influential media sites (LIM predicts well the diffusion of Memetracker phrases), the most active Twitter users have less influence on the overall adoption of hashtags. In addition, notice that α-LIM and B-LIM further increase the performance over the LIM on the Memetracker dataset. This means that the novelty of a phrase and imitation are important factors in the diffusion of textual phrases. On Twitter B-LIM slightly outperforms LIM, while α-LIM performs poorly, which hints that diffusion of hashtags is also driven by imitation, while recency does not play much role.

Table I also shows the performance of models based on the shape of the volume over time. Our previous research found that there are 6 distinct types of temporal variation in online media [2]. We cluster the volume curves into 6 clusters and Table I plots the temporal pattern of the centroid of each cluster. We note that AR and ARMA give even performance improvement over all types of volume curves, while the family of LIM models performs particularly well on phrases that exhibit a very abrupt spike in their volume. LIM can accurately model sudden spikes in adoption of textual phrases that are influenced by large media sites.

Analysis of influence functions. Our experiments so far demonstrated that LIM reasonably models information diffusion in online media. We now proceed to investigate how
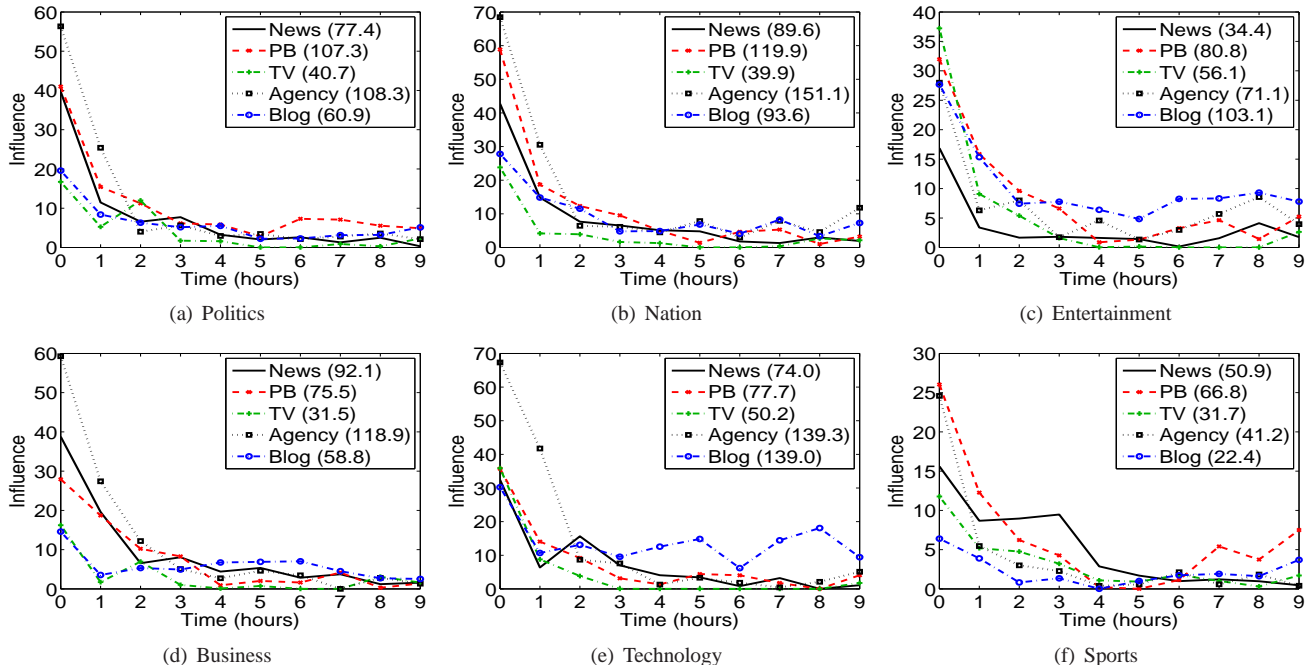
Figure 3. Average influence functions of five types of websites: Newspapers (News), Professional Blogs (PB), Television (TV), News Agencies (Agency), and Personal Blogs (Blogs). The number in brackets denotes the total influence of a media type.

the influence of various types of nodes changes depending on the topic of the information and the type of a node.

Memetracker dataset consists of a wide range of media sites from traditional mass media such as newspapers, nationwide TV stations and press agencies, to modern online independent news sites, professional and personal blogs. Since the credibility of information depends on the type of the source [21], we are interested in estimating the influence of a different types of media on the diffusion and adoption of textual phrases. Similarly to having different types of media, we also have different types of textual phrases. The intuition here is that different participants in online media discourse may have different influences depending on the topic of the debate [11]. In this respect we categorize textual phrases into six different topics. For each topic, we then estimate the influence functions of various types of sites (blogs, newspapers, etc.).

For the purpose of the experiment, we identify five types of media: Newspapers (New York Times, USA Today), Professional blogs (Salon, Huffingtonpost), TV stations (ABC, CBS), News agencies (AP, Reuters) and (personal) Blogs. In total we select 22 sites, and group them in the above five groups (the extended version of the paper [1] gives a full list). In order to find topics of textual phrases, we notice that several news sites specify the topic of an article in the URL. For each phrase, we simply list the URLs of all the articles that mention the phrase, and count which of the topic names (Politics, Nation, Entertainments, Business, Technology and Sports) appears in the URLs. When a single topic dominates, we consider the phrase to belong to that particular topic.

Now, we estimate the influence functions of 22 media sites ($N = 22$) on each of the six topics, by fitting LIM with the phrases in the topic. We plot the average influence function of sites of that particular type. Note that the influence functions model the influence per mention, whereas we are interested in the amount of total influence that each type of media has on the diffusion of phrases. In order to obtain the total influence, therefore, we normalize the influence functions of each type with the average number of the mentions of phrases on particular topic.

Figure 3 gives the influence functions for the five types of media and six topics. In the legend of the figure, we also compute the total influence of a media type by summing the values of their influence functions. Notice that in general influence functions tend to decay rapidly over time. While the decay is particularly pronounced for business and politics, for entertainment or sports the influence seems to last somewhat longer. Similarly, the influence of bloggers tends to be lower at start, but tends to last longer (in particular for entertainment and technology). This confirms the intuition that blogs tend to be echo chambers while mainstream media play the dominant force in the news cycle [23]. This is further confirmed by the fact that politics, business, technology and the nation tend to be dominated by news agencies. Professional blogs are the second in terms of total influence in politics and national news, newspapers are the second in business, and personal blogs are in technology. In entertainment and sports, the situation is somewhat reverse. For entertainment it is the personal blogs that are the most influential, while for sports it is the professional blogs

| (a) Politics | (b) Nation | (c) Entertainment |

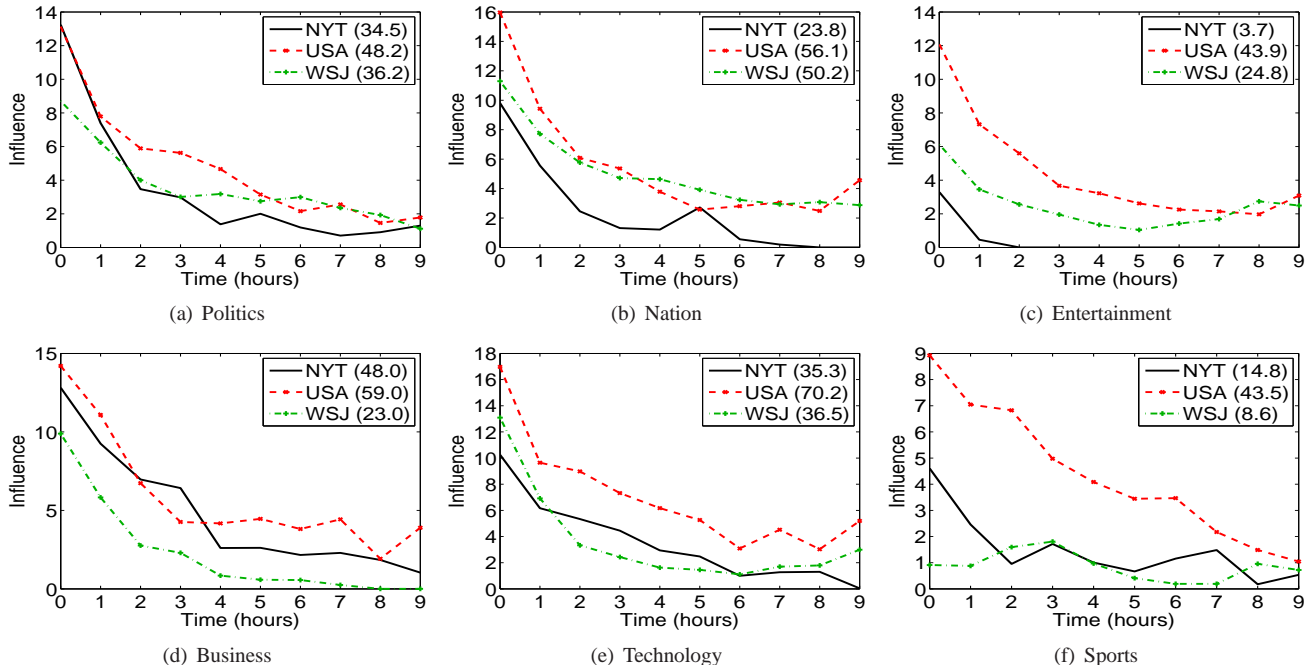| (d) Business | (e) Technology | (f) Sports |

Figure 4. Influence functions of the New York Times (NYT), the Wall Street Journal (WSJ), and USA Today (USA).

followed by the newspapers.

We repeat the same experiment with different setting where we model the total volume over time based only on three major U.S. newspapers: The New York Times (NYT), The Wall Street Journal (WSJ), and USA Today. Note that this model is particularly simplistic as it tries to model the diffusion of a textual phrase across the entire news media space based only on the information about three (i.e, $N = 3$) media sites. Figure 4 gives the influence functions for the three newspapers on the six topics. The USA Today is the most influential for sports and entertainment. However, we find the strong influence of the USA Today on technology somewhat surprising. While the New York Times has influence mostly in politics and business, the Wall Street Journal has more influence in national news, surprisingly in entertainment but not much in business.

All in all, these results agree with the intuition and are also consistent with the two-step flow model, coming from sociology and political science. Moreover, it is interesting that our model is able to detect and distinguish the fine differences between the roles that different types of media play in disseminating information of different topics.

Accounting for imitation. As we noted in the time series prediction task (Table I), the variants of the linear influence model that explicitly account for imitation (B-LIM) and recency (α-LIM) tend to perform slightly better than the straight LIM model. This is particularly the case in diffusion and adoption of textual phrases related to news, where imitation and recency play important roles.

We first explore the imitation. As before we take $K =$

1, 000 highest volume textual phrases and $N = 100$ websites that mentioned most of these phrases. We then fit the B-LIM model, and in Figure 5(a) we plot the latent volume $b(t)$ as a function of t. On the plot we also show the appropriately scaled average phrase volume, $\bar{V}(t) = (1/K) \sum_k V_k(t)$. Here, we index the time t so that the chronological median of the mentions of each phrase occurs at $t = 0$. Notice that the latent volume tightly follows the average volume over time, especially on the upward part. We also observe that the imitation effect reaches its maximum just before the phrase has its peak volume (i.e., $b(t)$ peaks just before $\bar{V}(t)$ does).

Given these results we also compute an average number of mentions of a phrase per website, $\bar{M}_u(t) = \sum_k M_{u,k}(t)$. We then find the media site with the highest correlation of the number of mentions $\bar{M}_u(t)$ with the imitation $b(t)$. We find that it is the Associated Press (AP) that best approximates the amount of imitation over time, $b(t)$. This confirms that articles that appear on AP are automatically distributed over hundreds of sites (that subscribe to AP's news feed) within a few hours.

Accounting for novelty. We also evaluate the effects of recency and novelty on the diffusion of textual phrases in online media. We fit α-LIM which estimates the recency factor $\alpha(t)$ as well as the individual influence functions.

Figure 5(b) plots the recency factor $\alpha(t)$ as a function of t. We observe some volatility in the recency factor long before its peak. Our intuition is that, in this period, the information is still developing with additional events, controversies and other external factors that make $\alpha(t)$ unpredictable. However, the rest of the recency factor $\alpha(t)$
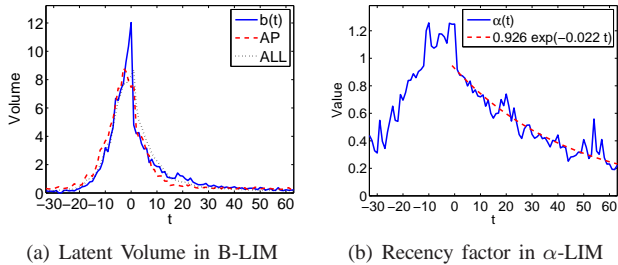
(a) Latent Volume in B-LIM   (b) Recency factor in α-LIM

Figure 5. Latent volume $b(t)$ and the recency factor $\alpha(t)$.

(a) Average influence functions of each group based on volumes   (b) Average influence functions of each group based on follower counts

Figure 6. Influence functions of groups of Twitter users.

can be very nicely explained. We notice that, for about ten hours, the effect of recency is the strongest and only later starts to slowly decay, with slower decay than the uptake.

To gain further insights into how effects of recency and novelty decay over time, we fit an exponential decaying function $\alpha(t) \approx ce^{-\lambda t}$, and find $c = 0.93$ and $\lambda = 0.0215$ to give the best fit. As shown in Figure 5(b), the exponential decay function very closely approximates $\alpha(t)$. Based on the value of decay parameter λ, we can estimate the half-life time τ such that $\alpha(\tau)$ is a half of $\alpha(0)$. We find the half-life to be 32.2 hours, which is about a day and a half and suggests that people consume news on daily basis.

Influence of users on Twitter. Last we explore the influence functions of Twitter users. Since the Twitter data is very sparse in a sense that each user mentions relatively few different tags, we consider a set 10,000 Twitter users, and aggregate them into 100 groups of 100 users. We consider two different types of grouping. First, we order users by the amount of their activity (hashtag volume) and second we order them based on the number of their followers. We fit B-LIM and examine the relation between the hashtag volume and the influence they have on the adoption of hashtags across the whole Twitter network.

Figure 6(a) shows the amount of influence of users grouped based on their total volume. All groups tend to have similar form of total influence. The group with the third largest volume has the most total influence, while the highest volume group has the lowest. Similarly, Figure 6(b) shows the influence functions of users grouped based on their total number of followers (i.e., in-degree) in the Twitter social network. Surprisingly, we find that the Twitter users with the intermediate number of followers have much higher influence than the highest in-degree nodes. While our results are somewhat different from literature in viral marketing and word of mouth [32], [25] which often assumes nodes with the highest follower count to be most influential, our results are consistent with the recent findings [7] which suggest that users with the highest follower count are not the most influential in terms of information diffusion. Rather, users with the number of followers of around 1,000 tend to be most effective in diffusion and adoption of hashtags.

The results on Twitter nicely align with the experiments on Memetracker data. As the Memetracker experiments
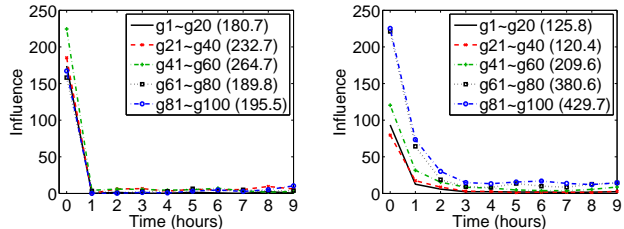
focused on online media and adoption of short textual phrases, we find that the mainstream media holds the most influential position in the dissemination of news content. On the other hand, hashtags on Twitter are a very different type of contagions. Hashtags are not news but rather socially contagious tags that are adopted in a distributed manner without a central supervision. Therefore, the diffusion of hashtags is mostly governed by the Twitter social/information network. This way Twitter users with "too high" number of followers, which usually correspond to celebrities and organizations, may be very influential in propagating the "information" contagions such as news, but not in diffusing more "social" contagions such as hashtags.

IV. CONCLUSION

We started with an assumption that the diffusion of information and other contagions is governed by the influence of individual nodes. Instead of focusing on the network topology and formulating a problem of predicting which node will infect which other individual nodes, we develop a Linear Influence model, where the influence functions of individual nodes govern the overall rate of diffusion through the network. We developed an efficient model parameter estimation method that is based on simple least squares-like formulation. Adopting a non-parametric modeling of the influence functions allowed us to accurately model and predict how diffusion unfolds over time.

We experimented with a set of 500 million tweets and a set of 170 million news media articles. Besides demonstrating that LIM outperforms classical time series prediction methods, we also gain a number of insights. For example, we identified influence functions of various websites and found that they heavily depend on the type of the website and the topic of the information. Furthermore, we also observed that the imitation and novelty have a strong force on the adoption of short textual phrases in online news media. As the adoption of short, news-related textual phrases appears to be highly governed by the influence of the few large media websites, the adoption of Twitter hashtags is governed by a much larger set of active users, each of which has relatively less influence. Moreover, we also observe that users with the most followers are not the most influential in propagating hashtags.

Our work opens up a new framework for the analysis of the dynamics of the information diffusion and influence in (implicit) social and information networks. Our models are broadly applicable to general diffusion process, as they do not require knowledge of the underlying network. An interesting venue for future work is to extend the model to allow for non-linear effects and to automatically discover the types of roles different participants have in the diffusion of information.

REFERENCES

[1] Supplementary material: `http://tinyurl.com/2dxltml`.

[2] J. Yang and J. Leskovec. Patterns of temporal variation in online media Technical Report, Stanford Infolab, 2010.

[3] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. In *Web Intelligence*, pages 207–214, 2005.

[4] N. T. J. Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Hafner Press, 2nd ed., 1975.

[5] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Prentice Hall, 1994.

[6] J. J. Brown and P. H. Reingen. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 14(3), 1987.

[7] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM '10*, 2010.

[8] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW '09*, 2009.

[9] T. F. Coleman and Y. Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM J. of Optimization*, 6(4), 1996.

[10] N. Friedkin. *A Structural Theory of Social Influence*. Cambridge University Press, 1998.

[11] K. E. Gill. How can we measure the influence of the blogosphere? *Workshop on the Weblogging Ecosystem*, 2004.

[12] M. Goetz, J. Leskovec, M. Mcglohon, and C. Faloutsos. Modeling blog dynamics. In *ICWSM*, 2009.

[13] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 3(12):211–223, 2001.

[14] A. Goyal, F. Bonchi, and L. Lakshmanan. Learning influence probabilities in social networks. In *WSDM '10*, 2010.

[15] M. S. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[16] J. Harsin. The rumour bomb: Theorising the convergence of new and old trends in mediated U.S. politics. *Southern Review: Communication, Politics and Culture*, 39(1), 2006.

[17] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[18] S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2):256–276, 2006.

[19] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

[20] M. O. Jackson and B. Golub. Naive learning in social networks: Convergence, influence and wisdom of crowds. *Working paper*, June 2007.

[21] T. J. Johnson and B. K. Kaye. Wag the blog: How reliance on traditional media and the internet influence credibility perceptions of weblogs among blog users. *Journalism & Mass Communication Quarterly*, 81(3):622–642, 2004.

[22] E. Katz and P. Lazarsfeld. *Personal influence: The part played by people in the flow of mass communications*. Free Press, 1955.

[23] B. Kovach and T. Rosenstiel. *Warp Speed: America in the Age of Mixed Media*. Century Foundation Press, 1999.

[24] C. L. Lawson and R. J. Hanson. *Solving least squares problems*. 3rd edition, 1995.

[25] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *EC '06*, 2006.

[26] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09*, 2009.

[27] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *SDM '07*, 2007.

[28] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *PNAS*, 105(12):4633–4638, 2008.

[29] E. M. Rogers. *Diffusion of Innovations*. Free Press, 1995.

[30] X. Song, Y. Chi, K. Hino, and B. L. Tseng. Information flow modeling based on diffusion rate for prediction and ranking. In *WWW '07*, 2007.

[31] D. J. Watts. A simple model of global cascades on random networks. *PNAS*, 99(9):4766–5771, 2002.

[32] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *J. of Consumer Research*, 34(4), 2007.

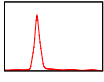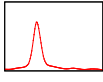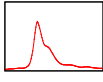[33] F. Wu and B. A. Huberman. Novelty and collective attention. *PNAS*, 104(45):17599–17601, 2007.
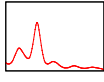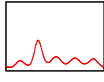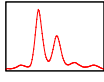
Cluster							ALL
AR	-3.14%	0.69%	1.13%	1.48%	0.76%	0.11%	-1.86%
ARMA	1.40%	0.11%	0.83%	2.65%	-0.12%	-0.09%	0.87%
LIM	15.16%	-25.50%	-19.03%	-15.47%	-18.50%	-9.64%	6.21%
B-LIM	15.36%	-25.74%	-19.08%	-15.38%	-18.12%	-10.77%	6.22%
ALIM	7.63%	-25.67%	-21.02%	-13.84%	-26.18%	-18.75%	3.53%
LIM + AR	-0.87%	-0.58%	0.91%	1.03%	0.75%	-1.54%	-2.41%

Table II

REDUCTION IN PREDICTION ERROR OVER 1-TIME LAG PERDICTOR ON TWITTER DATA FOR SIX TYPES OF SHAPES OF VOLUME OVER TIME. SEE THE MAIN TEXT FOR THE DESCRIPTION FOR MODELS.

Type	Website
Newspaper	nytimes.com online.wsj.com washingtonpost.com usatoday.com boston.com
Professional blog	huffingtonpost.com salon.com
TV	cbs.com abc.com
News Agency	reuters.com ap.org
Blogs	wikio.com forum.prisonplanet.com blog.taragana.com freerepublic.com gather.com blog.myspace.com leftword.blogdig.net bulletin.aarp.org forums.hannity.com wikio.co.uk instablogs.com

Table III
FIVE TYPES OF WEBSITES.

APPENDIX

Details for some websites in Table III. Counting the mentions from Associated Press (AP) is tricky as AP transmits its article to other websites before it posts on its site. We count the mentions from "breitbart.com" as the surrogate of the mentions from AP, because a mention from "breitbart.com" is the duplicate of an AP article in most cases, and it precedes other duplicates of the AP article. For the TV stations (ABC and CBS), we aggregate all the mentions from their local affiliates.