Seeing the Forest for the Trees: New Approaches to Forecasting Cascades

Siddharth Krishnan
Virginia Tech
siddkris@cs.vt.edu

Patrick Butler
Virginia Tech
pabutler@cs.vt.edu

Ravi Tandon
University of Arizona
tandonr@email.arizona.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Naren Ramakrishnan
Virginia Tech
naren@cs.vt.edu

ABSTRACT

Cascades are a popular construct to observe and study information propagation (or diffusion) in social media such as Twitter. and are defined using notions of influence, activity, or discourse commonality (e.g., hashtags). While these notions of cascades lead to different perspectives, primarily cascades are modeled as trees. We argue in this paper an alternative viewpoint of cascades as forests (of trees) which yields a richer vocabulary of features to understand information propagation. We develop a framework to extract forests and analyze their growth by studying their evolution at the tree-level and at the node-level. Moreover, we demonstrate how the structural features of forests, properties of the underlying network, and temporal features of the cascades provide significant predictive value in forecasting the future trajectory of both size and shape of forests. We observe that the forecasting performance increases with observations, that the temporal features are highly indicative of cascade size, and that the features extracted from the underlying connected graph best forecast the shape of the cascade.

CCS Concepts

•**Information systems** → **Web mining;** •**Computing methodologies** → *Machine learning;*

1. INTRODUCTION

A popular approach to studying social media chatter and information propagation in social networks is to characterize the occurrence and growth of cascades. Information cascades serve as a way of news and rumor spreading [15, 5], signal of online recruitment [8], a tool for viral marketing [13], etc. Given a medium such as Twitter, there are many notions of cascades, which afford varying levels of formal characterization and utility, e.g., retweet cascades, hashtag cascades, activity cascades, and URL cascades. They all differ in the underlying network over which diffusion is modeled, in the assumptions made to posit edges in the cascade, and the types of mathematical models that can support their rigorous analysis.

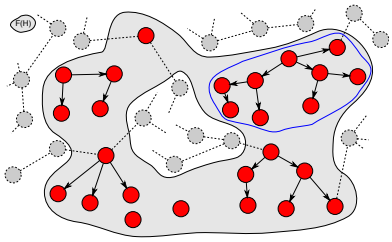

Figure 1: *A forest of diffusion trees*: The diffusion trees that propagate common information are of different shapes and sizes. We study the growth of the information cascade as a union of diffusion trees. The largest tree of the forest is outlined in blue.

Not withstanding the diversity of cascade notions available, the structure of cascades naturally lends itself to be viewed as a tree. In this work, we demonstrate that adopting a broader viewpoint of information cascades as forests (of trees) yields a richer vocabulary of features to understand information propagation. This viewpoint is advantageous for three reasons. First, most social media phenomena that go viral truly have multiple origins (or roots) rendering forests the more natural metaphor for modeling. For instance, a single hashtag is best modeled as multiple cascades rather than a single cascade. Second, information cascades, not withstanding their widespread use in research, seldom grow to significant sizes; modeling forests of cascades allows us to better 'pool' limited data and obtain greater specificity for classification and forecasting purposes. Third, forests enable the study of not only cascades' growth but also when new diffusion trees will emerge, thus providing a bigger picture viewpoint to the propagation of information in a network.

While the forest viewpoint offers the aforementioned advantages, it presents significant challenges from a forecasting perspective. While predicting the growth of a diffusion tree, as in [4], we are forecasting the *extension* of an observed structure. In the forest model, the growth is two-dimensional. Not only are we forecasting the growth of individual trees but we also forecast how new diffusion trees will emerge. We demonstrate how using the *underlying connected graph* [14] and other properties of the forest enable us to successfully tackle that challenge.

WebSci '16, May 22-25, 2016, Hannover, Germany

DOI: http://dx.doi.org/10.1145/2908131.2908155

Our key contributions here can be summed up by answering the following questions:

1. **How can we quantify the *size* and *shape* of an *information cascade forest?*** The forest model describing an information cascade has a two dimensional growth - *tree-level* and *node-level*. To account for both dimensions, we use three different measures—*largest tree size, forest size, and cascade size*. The Wiener Index (WI) [6] was shown to be indicative of a diffusion tree's structural virality. We extend WI in the context of a forest cascade. We also present another formulation to characterize structural virality, namely Virality Potential Index (VPI). VPI overcomes the shortcoming of WI by explicitly taking into account the underlying propagation network. In addition, VPI (explained in Sec. 4) has an interesting probabilistic interpretation: it can be viewed as the probability of all the inactive followers the users in a diffusion network becoming simultaneously active. A high value of VPI thus represents the *virality potential* of the diffusion network.

2. **Can we characterize an information cascade forest from a forecasting viewpoint?** The unpredictability of information cascades [21] and the rarity of large cascades [6] have been argued in literature. We tackle both problems by presenting a framework that extends the idea proposed by Cheng et al. [4] to forecast cascade forests as a series of prediction tasks. Our cascade model presents the unique opportunity to capture the relationship between diffusion trees propagating the same discourse commonality. We demonstrate how four classes of features—forest structural, underlying connected graph, Twitter follower features, and temporal aspects of forest growth—not only best describe the forest, but also provide significant predictive value in forecasting the cascade's future trajectory. We compare the performance of each class of features. Our experiments show that *temporal* features are best indicative of size and *underlying connected graph* features can best forecast structure.

3. **Can we forecast a cascade sans temporal features?** Temporal features of the cascade have been critical in forecasting cascades [4, 25]. Given the snapshot of a cascade graph, we show that by capturing the structural dynamics between the several diffusion trees, it is possible to forecast the eventual *size* and *shape* of an information cascade forest. Particularly, we demonstrate a process to extract an underlying subnetwork and its properties that enables this forecasting success.

2. RELATED WORK

Recently significant progress has been made in predicting the future trajectory of information cascades. We limit our survey to work most relevant to ours. This prediction problem has been formulated in terms of different figures of merit such as popularity [22] in Digg, re-tweeting on the Twitter network [18], user interests in microblogs [1, 20], and photo reshares [4]. Forecasting the volume or aggregate activity of a hashtag [16] or news phrases [26] has been tackled. Many studies approach the problem using a classic machine learning perspective such as regression [23, 12] or classification [10, 11]. Our work adds a new wrinkle to such prediction problems by studying the predictability of information cascades modeled as *forests of diffusion trees* that occur in social media. By extending the framework of analyzing the lifetime of a cascade as a series of prediction tasks proposed in [4], we show that not only are we able to successfully forecast the growth of the cascade forest, but also that this viewpoint presents a unique opportunity to engineer different categories of features that can describe a forest and capture the relationship between diffusion trees.

3. PRELIMINARIES

The mechanics of information cascade modeled as forest is presented in Fig. 1. We present a more detailed description below.

3.1 Diffusion Tree

The natural way of analyzing information propagation on Twitter is using *diffusion trees*, where each tree is a time-ordered sequence of connected nodes that mention a hashtag. Intuitively, these diffusion trees are constructed as follows. When a user posts a tweet, at time t, containing a hashtag and a few of his/her followers post a tweet containing the same hashtag at a later time, we add them to the tree and repeat this process till no more users can be added. More formally: let $G = (V, E)$ be a directed graph (the Twitter follower network in our case). We represent a tweet as $T(m, t)$, being a function of the user $m \in V$ and a timestamp t. Let the tweet $T(m, t)$ contain a hashtag H, i.e., $H \in T(m, t)$. We denote $\mathsf{Follow}(m)$ as the set of followers of user m. A diffusion tree $\mathsf{DiffTree}(m, H)$ as a function of the user m and hashtag H is then recursively defined as:

$$\mathsf{DiffTree}(m, H) = \{m\} \cup \{x \in \mathsf{DiffTree}(n, H) : n \in \mathsf{Follow}(m), H \in T(n, t^{'}), t^{'} > t\}$$

We always consider the first tweet by a user that mentions H. Our definition of activity propagation guarantees that the resulting diffusion structure is a tree since the influencer for a node entering a cascade is a neighbor who posted the latest tweet.

3.2 Cascade Forests

With the proliferation of hashtags across social media, characterizing their spread is more cogent with the notion of a *hashtag forest*, which can be defined as a collection of all diffusion trees that propagate H. More formally: Given a hashtag, H, a *forest* of H i.e. $F(H)$ is expressed as the union of all the *diffusion trees* that propagated the hashtag.

$$F(H) = \bigcup \mathsf{DiffTree}(u, H)\, \forall (u \in V) \wedge (\mathsf{DiffTree}(u, H)\, exists)$$

The main purport of this work is to analyze and forecast the future size and structural evolution of these information cascade forests.

3.3 Dateset Description

Our study primarily focuses on tweets spanning three major countries in the Latin American region: Brazil, Mexico, and Venezuela. Using the geocoder developed for EMBERS that uses the content of the tweet and poster's properties to estimate location of a tweet [19]. For our experiments, we harvested over ∼250 million tweets and culled cascades in a three month window (in Brazil from Jun. to Aug. 2013;

in Venezuela from Jan. to Mar. 2014; and in Mexico from Sep. to Nov. 2014) using hashtags that were widely propagating during that time period. To extract such hashtags, we ranked the ratio of each hashtag's three month count to their respective nine month count. Then we use a subgraph of the Twitter follower network (obtained through the Twitter API) for the Brazilian, Venezuelan and Mexican regions to extract information trees (as described above). The Twitter subnetwork for the three countries of interest consists of over 100 million nodes and over 2.2 billion edges. Finally, we group all the trees that belong to one hashtag to create the information cascade forest.

4. PROBLEM FORMULATION

Our characterization of information cascades using the notion of forests allows for the following quantities to measure it's *size* and *shape*:

1. ***Tree size***: The number of nodes in a given tree.
2. ***Forest size***: The number of trees in a given forest.
3. ***Cascade size***: The number of nodes in all the trees in the forest i.e. the sum of all tree sizes in a given forest.
4. ***Wiener Index*** of a diffusion tree.
5. ***Forest Wiener Index*** of an information cascade forest.
6. ***Virality Potential Index*** of an information cascade forest

Quantities (1)–(3) are self-explanatory; we discuss features (4)–(6) next.

Wiener Index and Forest Wiener Index

Recently, it was shown that the Wiener index (WI) is indicative of a diffusion tree's structural virality [6]. Analyzing the structural virality of a forest can be fairly complex as the growth of a forest is two dimensional, along the dimensions of *forest size* and *cascade size*. To account for both dimensions, we extend the Wiener Index (WI) in the context of forest growth by computing the average WI across all the trees in a forest. More formally, we define Forest Wiener Index (FWI) as follows. Let $\mathsf{dist(m,n)}$ be the length of the path between two nodes m & n in a *diffusion tree*. Given a hashtag H and a forest $F(H)$, for any tree Tree,

$$WI(\mathsf{Tree}) = \frac{\sum_{\forall (m,n) \in \mathsf{DiffTree}} \mathsf{dist(m,n)}}{|\mathsf{DiffTree}| * (|\mathsf{DiffTree}| - 1)} \; ;$$

$$FWI(F(H)) = \frac{\sum_{\forall \mathsf{Tree} \in F(H)} WI(\mathsf{DiffTree})}{|F(H)|}$$

The FWI captures the forest level *shape* in a quantifiable fashion. Since large trees are quite rare, computing either the median and maximum of the WIs in a forest skews the measure towards (resp.) too small or too big values. While FWI is coarse, it collectively captures the structural virality of an information cascade forest by accounting for each tree's structural virality in the forest.

Virality Potential Index of a Forest

The Wiener Index is limiting in capturing the potential of the diffusion tree as it does not account for the underlying network that acts as the medium for the propagation process. Therefore, to incorporate structure of the underlying propagation medium, we extend the notion of structural virality in the following manner. For each user m, consider the entire set of followers $\mathsf{Follow}(m)$. Now, consider a subset of $\mathsf{Follow}(m)$, namely the set of infected followers denoted by $\mathsf{Infected}(m)$ which consists of those followers who have tweeted the hashtag H. The ratio $r_m = \frac{|\mathsf{Infected}(m)|}{|\mathsf{Follow}(m)|} \in [0,1]$ can then be viewed as a the overall *infectability* power of a follower of this node. If each infected follower of this node were to infect its follower in an i.i.d. manner, the probability of *all nodes being successful* in infecting (i.e., reaching its maximal virality) would be $(r_m)^{|\mathsf{Infected}(m)|}$. Thus, the probability of *all nodes in a tree becoming simultaneously viral* would be

$$\prod_{m \in \mathsf{DiffTree}} (r_m)^{|\mathsf{Infected}(m)|} = \prod_{m \in \mathsf{DiffTree}} \left(\frac{|\mathsf{Infected}(m)|}{|\mathsf{Follow}(m)|} \right)^{|\mathsf{Infected}(m)|}$$

Since the above measure of virality is a probability $\in [0,1]$, we instead measure it on the log scale (for numerical stability) and define the Virality Potential Index (VPI) of a tree as:

$$\begin{aligned} \mathsf{VPI}(\mathsf{DiffTree}) &= -\log \left(\prod_{m \in \mathsf{DiffTree}} \left(\frac{|\mathsf{Infected}(m)|}{|\mathsf{Follow}(m)|} \right)^{|\mathsf{Infected}(m)|} \right) \\ &= \sum_{m \in \mathsf{DiffTree}} -|\mathsf{Infected}(m)| \log \left(\frac{|\mathsf{Infected}(m)|}{|\mathsf{Follow}(m)|} \right) \\ &= \sum_{m \in \mathsf{DiffTree}} \mathsf{VPI}(m) \end{aligned}$$

which is the sum of the virality potential indices of all nodes in a tree.

Remark: It is interesting to note that the term $\mathsf{VPI}(m)$ is a non-monotonic function of the number of infected nodes $|\mathsf{Infected}(m)|$. In particular, this quantity which captures the potential virality of node m satisfies the following properties:

- $\mathsf{VPI}(m) = 0$ when $|\mathsf{Infected}(m)| = 0$, since the node has remained completely inactive so far.
- $\mathsf{VPI}(m) = 0$ when $|\mathsf{Infected}(m)| = |\mathsf{Follow}(m)|$, since the node has already infected all its followers and has no more potential infective power.
- $\mathsf{VPI}(m)$ is a strictly concave function of $|\mathsf{Infected}(m)|$ and is maximized when $|\mathsf{Infected}(m)| = |\mathsf{Follow}(m)|/e$. That is, VPI for a node m first increases from 0 as the

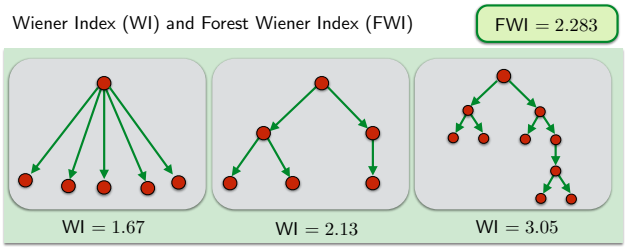

Figure 2: This figure shows a toy example of how Wiener index and Forest Wiener Index are computed. Second and third level diffusions lead to higher score for WI. The FWI captures the overall structural virality of the cascade forest in a quantifiable measure.

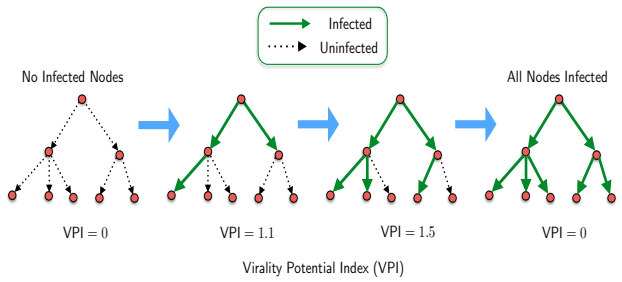

Figure 3: This figure shows the impact of infected nodes and the underlying follower graph on the temporal evolution of the Virality Potential Index (VPI).

number of infected nodes increase from 0 to $|\mathsf{Follow}(m)|/e$, and then gradually decreases back to 0 when the node has reached its full infective power, i.e., when $|\mathsf{Infected}(m)| = |\mathsf{Follow}(m)|$.

These properties are shown in Fig. 3, where the temporal evolution of VPI of a tree is shown as a function of the underlying follower graph and the infected nodes.

Using the definition of VPI(DiffTree), we then define the VPI of a forest $F(H)$ for a hashtag H as

$$\mathsf{VPI}(F(H)) = \frac{\sum_{\mathsf{DiffTree} \in F(H)} \mathsf{VPI}(\mathsf{DiffTree})}{|F(H)|}$$

The above measure rewards nodes for having more followers and infecting a good fraction of their followers while penalizing infected nodes that do not have followers and nodes that do have followers but cannot spread the infection. Moreover if two identical trees are part of a cascade, WI gives them both equal virality but the above measure distinguishes them based on their position in the network.

Formulating the Forecasting Problem

The rarity of large cascades has been well-documented in the information propagation literature [7] and formulating the cascade prediction question as a traditional regression or a classification problem raises issues of unbalanced classes and skewed predictions. Our dataset is no different. We demonstrate in Fig. 4 that each quantity that captures the size of the cascade has a skewed distribution. We observe that the curve for the forest size quite closely matches the cascade size. This conforms with the observation & prior research [10, 5] that there are many small trees in an information cascade and this is demonstrated by the power-law behavior of the tree size. A power-law curve fit to their tails, shows that the tree-size, forest-size and cascade-size have an α value of 1.85, 2.26 and 2.08 respectively. Similar to [4], we cast our prediction question in the form of a binary classification problem by asking whether a quantity of interest reaches the median. A *random guessing baseline* will thus have 50% accuracy. This problem is akin to asking if the size of the cascade will double using the following calculation. Since $\alpha \approx 2$,

$$\int_{x_{min}}^{f(x)} \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha} dx = \frac{1}{2} \implies f(x) = 2x_{min}$$

5. ATTRIBUTES AFFECTING INFORMATION CASCADE DYNAMICS

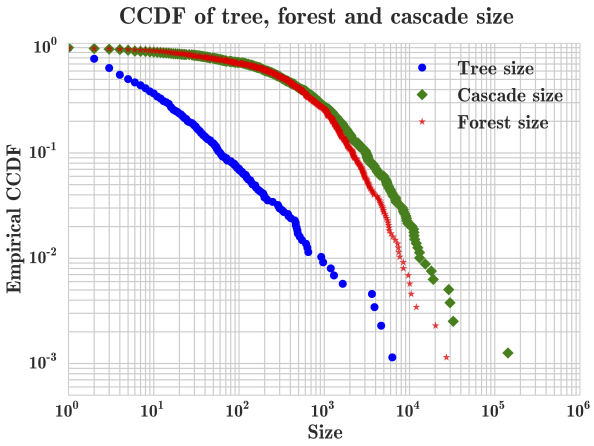

Figure 4: We plot the empirical complementary cumulative distribution function for the quantities of interest viz. *tree size*, *forest size*, and *cascade size*.

The underlying network structure and the temporal properties of the information that is propagating on the network are instrumental in predicting cascade growth [4, 12]. We primarily use attributes from these two classes to engineer features for our forecasting algorithm.

5.1 Features for Diffusion Tree Dynamics

We adapt the features presented in [4] to our context (more details in Table 1) of diffusion trees by grouping them into three classes – *structural*, *temporal*, and *root node* (features of the original poster). We briefly describe the intuition behind these features.

Structural: The twitter network, capturing the follower-friend relationship, acts as the medium for the flow of information and hence the structural features are indicative of the tree's potential. Virality and cross-community content spread have been predicted using properties of underlying network structure [24].

Temporal: The rate of cascade's growth has been instrumental in predicting its virality. Previous work on conversation lengths in Facebook [2], predicting content popularity [22] and the ability of a cascade diffusing in the network [25] have all emphasized the predictive value of temporal properties. We adapt some of the well-known properties of structural, temporal aspects, and original poster's attributes of information diffusion trees (Table 1) to forecast their growth of the tree.

Root node: We harvested features from the Twitter metadata of the original poster (root node) and inferred features such as gender, Twitter age, average post rate.

5.2 Features for Forest Dynamics

We present four categories of features that we use to forecast the growth of the information cascade forest viz. *forest structural, underlying connected graph, Twitter follower network,* and *forest temporal.* A comprehensive list of the features used is presented in Table 1.

Forest Structural: We extract some collective topological features of the forest for our forecasting algorithm. The

Table 1: The features used in forecasting the growth of diffusion trees (top three categories) and information cascade forests (bottom four categories).

		Diffusion Tree - Structural features
S.1	Degree	Total number of followers of the i^{th} node of a tree
S.2	Induced Degree	Number of followers of the i^{th} node among the first k nodes
S.3	Active Nodes	Number of total users in the underlying graph reachable from the first k nodes and the root
S.4	Original Conn.	Number of neighbors of the root node who are in the tree
S.5	Subgraph	Number of edges on the induced sub-graph of the first k nodes from the underlying network
S.6	Border Nodes	Number of neighbors of the participating nodes that are not part of the tree
S.7	User Left	Number of nodes in the tree are not neighbors of the root
		Diffusion Tree - Temporal features
T.1	# Views	Number of users who saw the posts by the first k nodes of the tree
T.2	Avg. First	Average time between posts of the first $k/2$ nodes
T.3	Avg. Last	Average time between posts of last $k/2$ nodes
T.4	Elapsed Time	Time elapsed between the post by the root node and the post by the k^{th} node
T.5	Growth Rate	Change in time between posts of successive nodes in the tree
		Diffusion Tree - Root node features
R.1	Infectivity	Total number of root's followers infected at the k^{th} reshare
R.2	Post count	Number of tweets with hashtag of interest posted
R.3	Post count	Out degree of the root
R.4	Retweets	Number of retweets of the original post
R.5	Twitter Klout score	An influence score obtained from the Twitter metadata
R.6	Avg. Post Rate	Number of tweets posted per day by the root
R.7	Gender	Gender of the root poster, if available (categorical variable of three categories)
R.8	Twitter Age	Time since active on Twitter
		Forest - Structural G_f features
F.1	G_fEdges	Number of edges in G_f
Forest F.2	G_fTrees	Number of individual trees in G_f
F.3	G_fLargestTree	Size of the largest tree in G_f
F.4	G_fIsoNodes	Number of isolated nodes (trees with size = 1)
F.5	G_fDeg1Nodes	Number nodes with $outdeg = 1$
F.6	G_fSize3Trees	Number of trees with size greater than 3
F.7	G_fBroadcast	Number of broadcast trees (depth = 1)
F.8	G_fDensity	density of G_f, given as $\frac{G_f\text{Edges}}{\lvert N_f\rvert * \lvert N_f\rvert - 1}$, where N_f is number of nodes in G_f
F.9	G_fLargestTreeProp	The ratio between size of the largest tree and the number of nodes in the forest, essentially $\frac{G_f\text{LargestTree}}{\lvert N_f\rvert}$
		Forest - Underlying connected graph G_{uc} features
UC.1	G_{uc}Edges	Number of edges in G_{uc}
UC.2	G_{uc}ConnNodes	Number of connector nodes in G_{uc}
UC.3	G_{uc}ConnEdges	Number of edges incident to connector nodes
UC.4	Node Ratio	Ratio of number of nodes in G_f to G_{uc} (N_f/N_{uc})
UC.5	Edge Ratio	Ratio of number of edges in G_f to G_{uc} (G_fEdges/G_{uc}Edges)
UC.5	G_{uc}InDeg2	Number of connector nodes with $indeg \geq 2$
UC.6	G_{uc}Triads	Number of triangles in G_{uc}
UC.7	Tree Betweenness	mean of shortest paths in G_{uc} between trees in G_f
UC.8	G_{uc}Density	density of G_f, given as $\frac{G_{uc}\text{Edges}}{\lvert N_c\rvert * \lvert N_c\rvert - 1}$, where N_c is number of nodes in G_{uc}
		Forest - Twitter follower network G_{fn} features
FN.1	Root Followers	Number of Twitter followers of the root nodes of *all trees* in G_f
FN.2	Root Influence	Neighbors of the root in the G_f divided by the number of followers of the root node
FN.3	Leaf Influence	Number of followers of the leaf nodes of all the trees
FN.4	Forest Potential	Total number of followers of all the nodes in all the trees
FN.5	Follower Influence	$\forall v \in G_f$, the ratio of followers of v in G_f to all the followers of v in G_{fn}
FN.6	Border followers	Number of followers of nodes in G_f, that are not in G_f
		Forest - Temporal features
TF.1	Total Time Elapsed	Time elapsed between post of the first node in G_f and ith node in G_f
TF.2	Node Growth Speed	Average time between nodes for the i tweets
TF.3	Tree Growth Speed	Average time between trees for the m trees arising out of the i nodes
TF.4	Cascade Growth Rate	Change in times between successive nodes appearing in the forest
TF.5	Forest Growth Rate	Change in times between successive trees appearing in the forest
TF.6	Time Elapsed Origin	Change in times between first node's post and the ith node's post

number of trees that make up a forest and the distribution of their sizes provides information about the structure of the forest. Recently it was shown that there are two types of diffusion trees - *broadcast* and *multi-level diffusion* [6] (see Fig.5e) and a correlation was found between virality and the type of the tree. The second level cascading process correlated highly with virality. We track the type of the tree in our feature list. We capture the imprint of the largest tree in the forest by computing the ratio between the size of the largest tree and that of the information cascade forest.

Underlying Connected Graph (UCG): Our forest is a collection of trees, which are essentially disconnected components in a large graph i.e. the Twitter follower network. We adapt the idea presented in [14] and use their algorithm to build an underlying connected graph which combines all the trees into one connected component. We introduce connector nodes C from the Twitter follower network into the forest of trees to achieve this. We find the set of connector nodes between the different trees via shortest paths and use the least number of nodes that makes the forest into one connected component (G_{uc}). We then extract several graphical features of G_{uc}. More formally the underlying connection graph is as follows:

Given $F(H) = G_f(V_f, E_f)$, $G_{fn}(V_{fn}, E_{fn})$ is the Twitter follower network, connector nodes $C \subseteq V_{fn}$, then the underlying connected graph $G_{uc}(V_{uc}, E_{uc})$ is built such that: $V_{uc} = V_f \cup C$ and $E_{uc} = \{(u,v) \mid u \in V_f, v \in C \, and \, (u,v) \in E_{fn}\}$

Finding an underlying connected graph is NP-hard as one can reduce the Steiner tree computation to this problem. We use the heuristic proposed [14] to construct our underlying connected graph. Briefly the method is as follows. Consider the various trees in the forest and sort them in descending order of size. Connect the largest tree to the second largest tree via the shortest path between them to make it one component, and then proceed to connect trees to this component in decreasing order of tree size.

The underlying connected graph captures the structural relationship between the various trees in the forest. When a hashtag has several entry points into the forest, it is quite natural that they can be spread apart initially and become dense as the forest grows. The density of the G_{uc} gives us this information. Also, we measure tree betweenness as the average of all the shortest paths between each pair of diffusion trees. This quantifies how far spread out the forest is. Finally, the connector nodes with high in-degree are quite likely to become part of the cascade during its growth following the intuition proposed by several threshold models [9, 17]. The features extracted from the G_{uc} have high forecasting potential, especially in forecasting the structural growth of the cascade.

Twitter Follower: The overall influence of the trees in the information cascade and the individual influence of the nodes is best captured by the footprint of the cascade on the Twitter follower network. As shown in [11, 18] features extracted from the follower network, such as number of followers of nodes in the cascade, are very useful in predicting cascades. We capture this phenomenon at various levels using *root followers*, *forest potential*, and *leaf influence*. We compute features that capture the relationship between the cascade forest and the follower network. In particular, we capture the influence each node has using ratios computed between nodes in the cascade and the nodes not in the cascade. We also extrapolate some of the tree-level features such as *border nodes* in the context of forests.

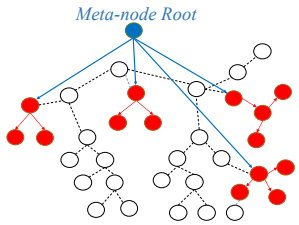

Figure 6: We add a *meta-node* and make it the root of all the trees in a cascade forest. We extract the tree level features of this artificially extended tree and use it as a baseline for our approach.

Forest temporal: The significance of temporal properties of cascades underscored in earlier work [2, 22, 25] coupled with the forecasting potential demonstrated in [4], we extend some of the temporal features to our cascade forests. We observe that this category of features have significant potential in forecasting cascades.

6. EXPERIMENTS

We demonstrate the performance of our forecasting algorithm both at the tree level and collectively at the forest level. Our feature set and forecasting tasks naturally lend themselves to a setup where we can ask the following questions:

1. Given that tree level forecasting models have been well studied, can we design a forecasting model using a tree based approach for the information cascade forest? (Sec. 6.1)
2. Which group of features has the most predictive power? (Sec. 6.2.1, 6.2.2, 6.2.3)
3. How does the performance of prediction vary with number of observations? (Sec. 6.2.1, 6.2.2, 6.2.3)
4. How early can we forecast the future size of the cascade? (Sec. 6.2.4)
5. Can structural virality and virality potential be forecasted? (Sec. 6.3)

Our experimental framework is setup as a series of prediction tasks, where we observe k nodes in an information cascade and forecast if a quantity of interest will reach the median or not. In section 4, we demonstrated that random guessing (one baseline), is always 50%. We describe a tree based approach in this section, which we use as another baseline. We experimented with different classifiers viz. SVM, decision tree, random forest, and logistic regression and the forecasts were similar across all of them (within ∼2%). The results presented are from the logistic regression based classifier. In all cases we performed 10-fold cross-validation. While all three evaluation metrics viz. precision, recall and F1-score obtained similar results, for ease of comparison we report the F1-score for all our experiments.

6.1 Meta-node Root Approach

We add a *meta-node* (shown in Fig. 6) to the cascade and make all the diffusion trees in a forest its first level children. Now the information cascade is a *tree*, albeit artificially. We extract the features (sans the artificially introduced root-node features) of this propagation tree (outlined earlier and

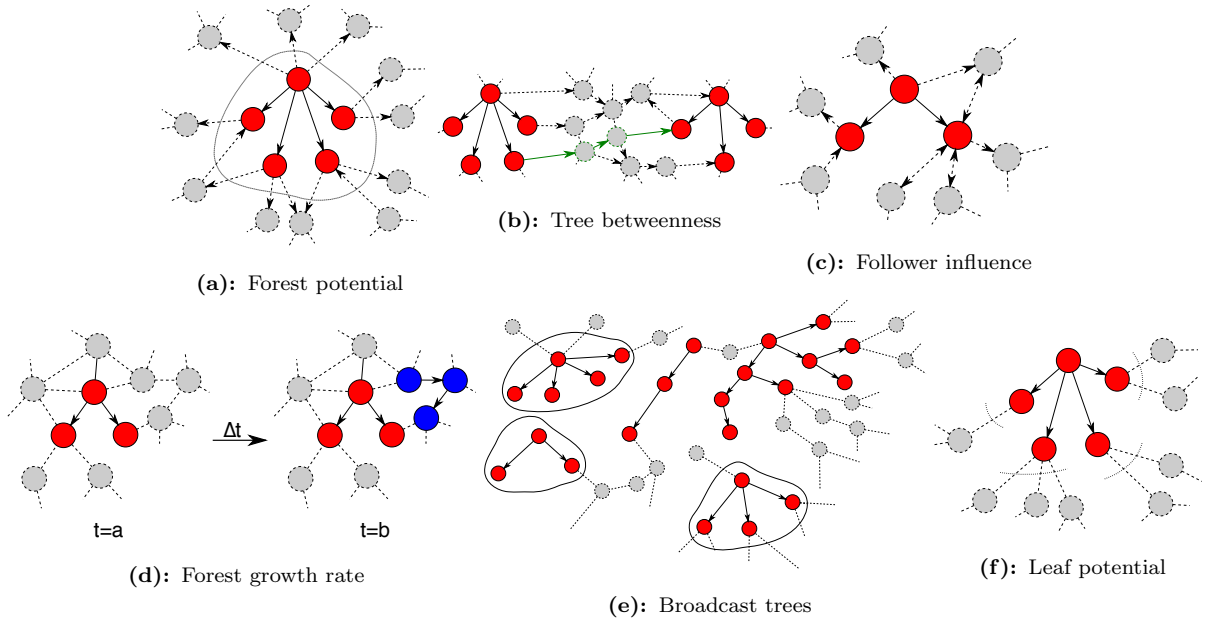

Figure 5: We pictorially show some of the forest level features that we extract from our data. Particularly, we use features from the *underlying connected graph* to forecast forest growth. We also extract several temporal signatures of the cascade forest.

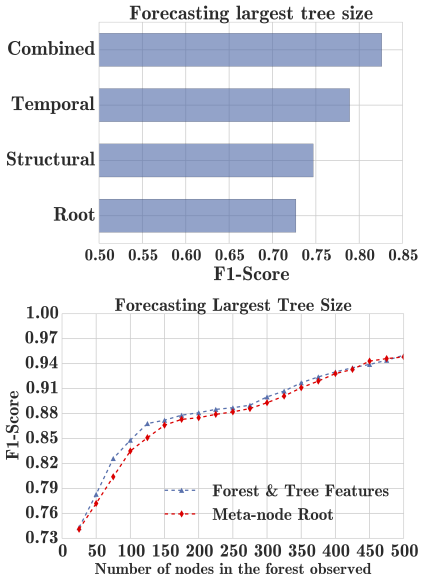

Figure 7: Forecasting the size of the largest tree: *Above*: We see that we are able to significantly outperform the *random-guessing* baseline. *Below:* Like the Facebook photograph reshare cascades observed in [4], the prediction accuracy improves with observations. Moreover, the meta-node approach which is a tree based method performs just as well as the forest features.

in Table 1), in our forecasting experiments. We use this approach as a baseline to evaluate the performance of the newly engineered forest features in our present work.

6.2 Forecasting Size

Among the quantities of interest defined earlier, *tree size, forest size,* and *cascade size* are related to size.

6.2.1 *Tree Size*

For this task, we forecast the growth of the largest diffusion tree in the forest. We see (Fig. 7: top) that, using the features extracted, our predictor performs up to ∼33% (if we observe 100 nodes in the forest) over the *random-guessing* baseline. To evaluate the contribution of each feature group, we trained a classifier using structural and temporal features separately. We observe that the temporal features (28% above baseline) outperform the structural features (23% above baseline) by about 5%. For the second experiment (Fig. 7 below), we evaluate the prediction accuracy with varying number of observations. We find that the forecasting accuracy improves with more observations of the forest, achieving close to 95% accuracy. Since this task is tree-level forecasting and the meta-node approach is tree-based, it is quite natural that its prediction accuracy is quite optimal. This result also conforms with the discovery in [4] that temporal features have the most predictive power and that accuracy increases with number of observations.

6.2.2 *Forest Size*

The forest model affords for a two-dimensional growth of information cascades. In this subsection, we focus on growth as a function of trees; more precisely we look at *forest size*. We observe an accuracy of 83% in forecasting the forest size (see Fig. 8 above). Temporal features obtain a performance of 78%. In the absence of temporal features of the cascade, with a snapshot of the cascade graph, the G_{uc} set of features predicts the growth of forests with 75% accuracy. This result

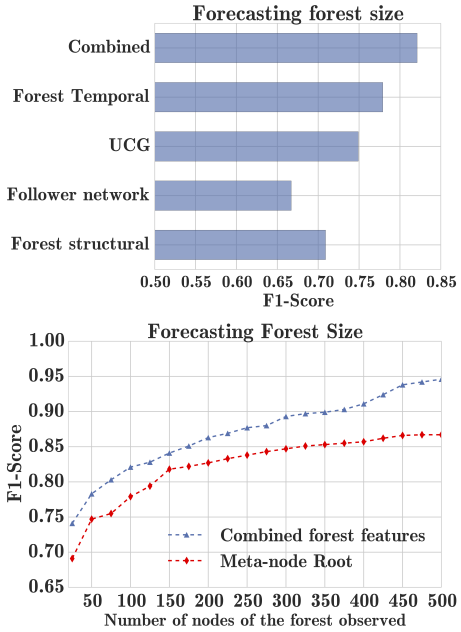

Figure 8: We are forecasting the *outbreak* of new trees in the forest using features from trees already in the forest, thus making this a very interesting result. The **UCG** is a very useful feature set as it uses a snapshot of the cascade graph to predict the formation of new trees. The temporal features by themselves are highly indicative of **forest-size**. We find that there is consistent improvement to the forecasts with increased observations. The set of forest features consistently outperform the meta-node based tree approach

is particularly useful because of the limitations posed by the Twitter API to extract the follower network. We observe that the combined forest features show an improvement of upto ∼9% over the meta-node tree based approach.

6.2.3 *Cascade Size*

We find that the classifier achieves around 82% performance (and upto ∼95% for large cascades of size > 500) when we combine all the features, when predicting cascade size. Temporal features are the best performing set in this case as well. We note that when the prediction question is cast at the node-level, the follower network features perform better than the structural features of the forest. This result agrees with intuition as this forecasting question looks simultaneously at the growth of a tree and the forest. Since the border nodes, induced degree, original connection and other tree level structural features are analogous to some of the features in the group of follower network features, and are excellent at predicting tree level growth, we find that follower network features are very good at predicting node-level growth of the forest.

6.2.4 *Fixed Target Cascade Size*

In the effort of finding a sweet-spot to forecast the eventual size of a cascade, we fix the lower-bound of cascade size and analyze how early we can predict it. Towards that end, we extracted a subset of all the forests with size of at least R and for varying k ($k \leq R$) observations, we forecast whether the cascade will reach the median (of all cascades of size at least R). We find that (Fig. 9) the performance increases consistently. In other words, there is no fixed number of observations after which the performance either remains the same or decreases. Although it is worth noting that the rate at which the performance improves is much higher with lower number of increments in nodes observed than with increments close to R. Each feature category's performance, as expected, improves with increased number of observations. We find the class of temporal features performs very well on its own. The forest's topological features' performance is the least with lesser observations but eventually outperforms follower network features with the luxury of observations. We find that the performance of underlying connected graph improves consistently with observations and performs almost as good as the temporal features.

6.3 Forecasting Structure

6.3.1 *Tree Structure*

In this set of experiments we forecast if the largest tree's Wiener Index (WI) achieves median value across all the forests' largest trees. Our classifier has a prediction accuracy of about 73% and concurring with the findings in [4], we find that the temporal features and structural features have similar performance.

6.3.2 *Forest Structure*

In forecasting the FWI, the features extracted at the forest-level obtain a performance of about 74%. While the forest-temporal features perform quite similar to the forest-structural features, we find that the G_{uc} set outperforms every other group of features significantly. Since FWI is a structural characterization, it is natural that it is highly correlated with structural properties of the forest. To understand this performance further, we trained a classifier individually on each of the features in the UCG. The *density* of G_{uc}, *InDeg2* and *ConnEdges* have the highest performance at ∼63% for each of them. The connector nodes fall on paths that connect the various information trees and are prime candidates to become part of the information cascade in its progression. Also, the connected edges in turn become part of the forest and contribute towards FWI. Therefore, it stands to reason that the features associated with those nodes and the underlying connected graph in general well capture the structural evolution of the forest.

Similarly, while forecasting VPI, we find that the structural features have the most forecasting potential. Since the VPI is characterized at the node level and then extended to the tree, it is not surprising that the *metanode* based approach forecasts the structural virality almost as well as all the forest features combined. Although, when the size of the cascade grows the forest features outperform the tree based approach. We also note that the G_{uc} features are best indicative of the VPI. This result is natural since VPI's characterization comes from the Twitter follower network.

6.4 A Viral Forest Cascade Example

In July 2013 MTV ran the *#mtvhottest* campaign for electing the most popular summer music star via Twitter. The campaign, started July 20th and ended August 18th, received over 320 million tweets worldwide and it was viral in our dataset as well with ∼ 1.98 million tweets. In the first twelve hours of the first appearance of the hashtag, the cascade garnered 32,584 tweets which made it one of the fastest growing cascades in our dataset. The cascade steadily grew over the duration of the next month collecting around 20-40k tweets on a daily basis. By the last

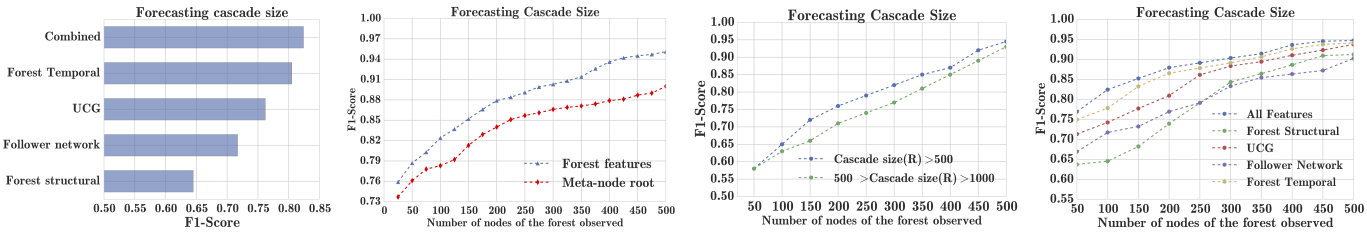

Figure 9: The follower network features are better than the structural features in forecasting *cascade-size* but have the least performance among all features while forecasting *forest-size*. While performance is quite different for each class of features in the early stages, with more observations, all the classes of features perform quite similarly. (right)

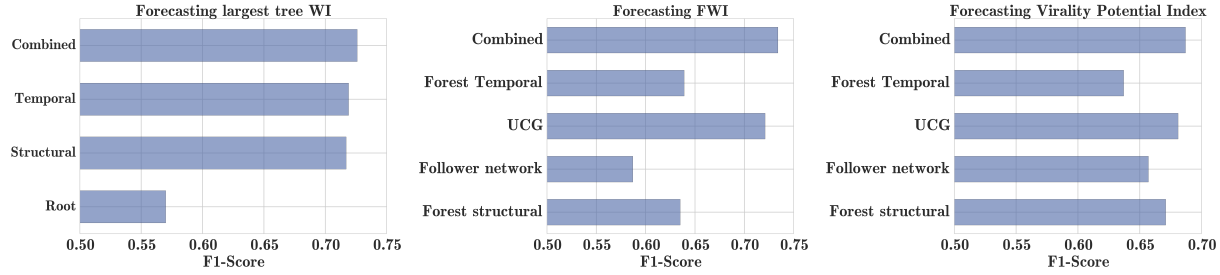

Figure 10: Forecasting structural virality of the largest tree and the entire forest. The underlying connected graph best forecasts the structure. The VPI is designed at the tree level and extended to the forest. Thus it is quite intuitive that the *meta-node* approach forecasts almost as good as the forest features.

week of the campaign, the cascade garnered close to 50k tweets per day. We observed that along with *#mtvhottest* other hashtags like *#JustinBeiber, #MileyCyrus, #OneDirection, #VivaPerry, #LadyGaga, #Gagalicious, #VamosBieber* (shown as tagclouds evolving over time in Fig. 11) appeared. Interestingly, the use of several hashtags led to the development of several sub-forests. Moreover we also observed that several users were canvassing for their favorite stars and attempting to convince their followers via the use of mentions. Our forecasting algorithm was able to determine the growth of *#mtvhottest* with a confidence of up to 95%.

7. DISCUSSION & FUTURE WORK

This work offers an alternate viewpoint to information cascades modeled as forests of diffusion trees and analyzes their predictability. Although each of the the four classes of features extracted viz. *forest structural, underlying connected graph, Twitter follower* and *forest temporal* have individual forecasting potential, the forest temporal feature set is the most indicative of the eventual size of the forest. In the absence of temporal features, given a snapshot of a cascade graph, the set of underlying connected graph (G_{uc}) features can not only forecast the eventual size of the cascades, but is the best indicator of *shape* (structural virality). Even though several diffusion trees are disconnected components, we were always able to extract a G_{uc} indicating that the Twitter follower network is a dense connected network. We find that the performance of prediction consistently improves with number of observations and that there is no 'sweet spot' after which the performance tapers.

While we have primarily analyzed some of the fundamental structural and temporal aspects of these forests to engineer features, one could extract rich features based on content, communities in the network, and user interest groups that can potentially improve the overall forecasting. Moreover, the forest model naturally lends itself to studying the propagation of URL cascades on social networks. We will adapt our model to study that as well. The interplay between the content of the URL and other features combined with the forest-level features can give us a richer feature set to analyze and forecast information cascades. The proliferation of hashtags also allows for extending the forest model to consider trees formed across multiple social networks. This will not only help us better characterize information flow but also analyze and understand propagation patterns across social networks. The forest model can capture co-occurring hashtags and URLs to build forests that combine two different types of diffusion trees and capture richer information propagation patterns.

Twitter activity has been often precursor of events such as protests [19] and activity cascades have been used in protest recruitment and protest out break predictions effectively [3]. In the future, we aim to extend our model to monitor event-based cascading activity to examine event forecasting potential of Twitter cascades for events like flu outbreak, elections, agenda setting, or civil unrest movements.

Acknowledgements: Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

8. REFERENCES

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proc. ACM*

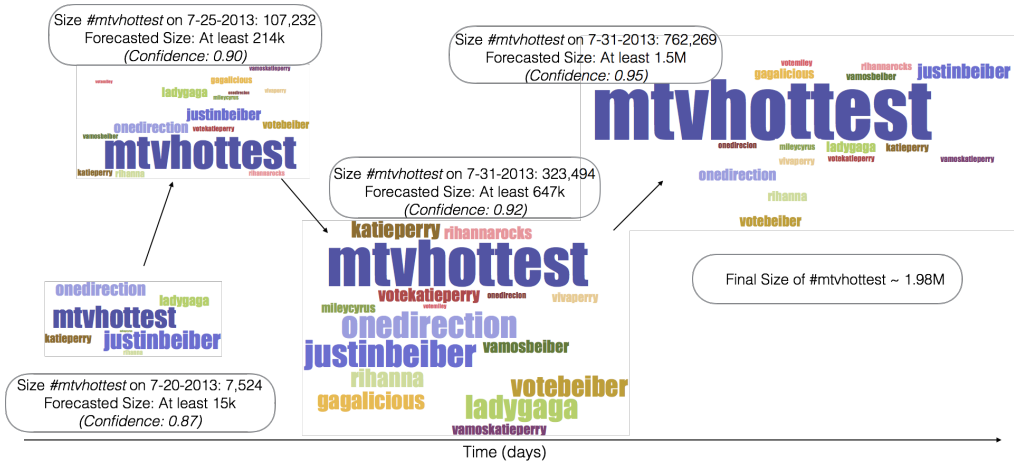

Figure 11: Illustration of cascade forecasting with forest features: *#mtvhottest* went viral as soon as it hit Twitter late July 2013. Several hashtags co-occurred with it. We show tagclouds of temporally evolving hashtags that were tweeted along with *mtvhottest*. Since the tagclouds are scaled, we observe that while *#mtvhottest* grows at a rapid rate the other hashtags, for example *#justinbeiber*, become smaller even though those cascades are evolving. Note that our confidence score keeps improving with more observations.

KDD, 2006.

[2] L. Backstrom, J. Kleinberg, L. Lee, and C. Danescu-Niculescu-Mizil. Characterizing and curating conversation threads: Expansion, focus, volume, re-entry. In *Proc. ACM WSDM*, 2013.

[3] J. Cadena, G. Korkmaz, C. J. Kuhlman, A. Marathe, N. Ramakrishnan, and A. Vullikanti. Forecasting social unrest using activity cascades. *PLoS ONE*, 10(6):1–27, 06 2015.

[4] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proc. WWW*, 2014.

[5] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In *Proc. ICWSM*, 2014.

[6] S. Goel, A. Anderson, J. Hoffman, and D. Watts. The structural virality of online diffusion. In *Mangement Science*, 2015.

[7] S. Goel, D. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *Proc. ACM EC*, 2012.

[8] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno. The Dynamics of Protest Recruitment through an Online Network. *Nature Scientific Reports*, 1, 2011.

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

[10] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proc. of WWW*, 2011.

[11] M. Jenders, G. Kasneci, and F. Naumann. Analyzing and predicting viral tweets. In *Proc. of WWW*, 2013.

[12] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev. Prediction of retweet cascade size over time. In *Proc. CIKM*, 2012.

[13] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007.

[14] J. Leskovec, S. Dumais, and E. Horvitz. Web projections: Learning from contextual subgraphs of the web. In *Proc. WWW*, 2007.

[15] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model. In *Proc. SDM*, 2007.

[16] Z. Ma, A. Sun, and G. Cong. On predicting the popularity of newly emerging hashtags in twitter. *JASIST*, 64(7):1399–1410, 2013.

[17] N. Pathak, A. Banerjee, and J. Srivastava. A generalized linear threshold model for multiple cascades. In *Proc. of ICDM*, 2010.

[18] S. Petrovic, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. In *Proc. ICWSM*, 2011.

[19] N. Ramakrishnan and et. al. Beating the news with embers: Forecasting civil unrest using open source indicators. In *Proc. KDD*, pages 1799–1808, 2014.

[20] D. Romero, C. Tan, and J. Ugander. On the interplay between social and topical structure. In *Proc. ICWSM*, 2013.

[21] M. Salganik, P. Dodds, and D. Watts. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311(5762):854–856, 2006.

[22] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Comm. of the ACM*, 53(8):80–88, Aug. 2010.

[23] O. Tsur and A. Rappoport. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *Proc. WSDM*, 2012.

[24] L. Weng, F. Menczer, and Y.-Y. Ahn. Virality prediction and community structure in social networks. *Sci. Rep.*, 3(2522), 2013.

[25] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in Twitter. In *Proc. ICWSM*, 2010.

[26] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *Proc. ICDM*, pages 599–608, 2010.