Can cascades be predicted?

Justin Cheng
Stanford University
jcccf@cs.stanford.edu

Lada A. Adamic
Facebook
ladamic@fb.com

P. Alex Dow
Facebook
adow@fb.com

Jon Kleinberg
Cornell University
kleinber@cs.cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

On many social networking web sites such as Facebook and Twitter, resharing or reposting functionality allows users to share others' content with their own friends or followers. As content is reshared from user to user, large cascades of reshares can form. While a growing body of research has focused on analyzing and characterizing such cascades, a recent, parallel line of work has argued that the future trajectory of a cascade may be inherently unpredictable. In this work, we develop a framework for addressing cascade prediction problems. On a large sample of photo reshare cascades on Facebook, we find strong performance in predicting whether a cascade will continue to grow in the future. We find that the relative growth of a cascade becomes more predictable as we observe more of its reshares, that temporal and structural features are key predictors of cascade size, and that initially, breadth, rather than depth in a cascade is a better indicator of larger cascades. This prediction performance is robust in the sense that multiple distinct classes of features all achieve similar performance. We also discover that temporal features are predictive of a cascade's eventual shape. Observing independent cascades of the same content, we find that while these cascades differ greatly in size, we are still able to predict which ends up the largest.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database applications—*Data mining*
General Terms: Experimentation, Measurement.
Keywords: Information diffusion, cascade prediction, contagion.

1. INTRODUCTION

The sharing of content through social networks has become an important mechanism by which people discover and consume information online. In certain instances, a photo, link, or other piece of information may get *reshared* multiple times: a user shares the content with her set of friends, several of these friends share it with their respective sets of friends, and a *cascade* of resharing can develop, potentially reaching a large number of people. Such cascades have been identified in settings including blogging [1, 13, 21], e-mail [12, 22], product recommendation [20], and social sites such as Facebook and Twitter [9, 18]. A growing body of research

has focused on characterizing cascades in these domains, including their structural properties and their content.

In parallel to these investigations, there has been a recent line of work adding notes of caution to the study of cascades. These cautionary notes fall into two main genres: first, that large cascades are rare [11]; and second, that the eventual scope of a cascade may be an inherently unpredictable property [28, 31]. The first concern — that large cascades are rare — is a widespread property that has been observed quantitatively in many systems where information is shared. The second concern is arguably more striking, but also much harder to verify quantitatively: to what extent is the future trajectory of a cascade predictable; and which features, if any, are most useful for this prediction task?

Part of the challenge in approaching this prediction question is that the most direct ways of formulating it do not fully address the two concerns above. Specifically, if we are presented with a short initial portion of a cascade and asked to estimate its final size, then we are faced with a pathological prediction task, since almost all cascades are small. Alternately, if we radically overrepresent large cascades in our sample, we end up studying an artificial setting that does not resemble how cascades are encountered in practice. A set of recent initial studies have undertaken versions of cascade prediction despite these difficulties [19, 23, 26, 29], but to some extent they are inherent in these problem formulations.

These challenges reinforce the fact that finding a robust way to formulate the problem of cascade prediction remains an open problem. And because it is open, we are missing a way to obtain a deeper, more fundamental understanding of the predictability of cascades. How should we set up the question so that it becomes possible to address these issues directly, and engage more deeply with arguments about whether cascades might, in the end, be inherently unpredictable?

The present work: Cascade growth prediction. In this paper, we propose a new approach to the prediction of cascades, and show that it leads to strong and robust prediction results. We are motivated by a view of cascades as complex dynamic objects that pass through successive stages as they grow. Rather than thinking of a cascade as something whose final endpoint should be predicted from its initial conditions, we think of it as something that should be *tracked* over time, via a sequence of prediction problems in which we are constantly seeking to estimate the cascade's next stage from its current one.

What would it mean to predict the "next stage" of a cascade? If we think about all cascades that reach size k, there is a distribution of eventual sizes that these cascades will reach. Then the distribution of cascade sizes has a median value $f(k) \geq k$. This number $f(k)$ is thus the "typical" final size for cascades that reached size

at least k. Hence, the most basic way to ask about a cascade's next stage of growth, given that it currently has size k, is to ask whether it reaches size $f(k)$.

We therefore propose the following *cascade growth prediction problem*: given a cascade that currently has size k, predict whether it grow beyond the median size $f(k)$. (As we show later, the prediction problem is equivalent to asking: given a cascade of size k, will the cascade double its size and reach at least $2k$ nodes?) This implicitly defines a family of prediction problems, one for each k. We can thus ask how cascade predictability behaves as we sweep over larger and larger values of k. (There are natural variants and generalizations in which we ask about reaching target sizes other than the median $f(k)$.) This problem formulation has a number of strong advantages over standard ways of trying to define cascade prediction. First, it leads to a prediction problem in which the classes are balanced, rather than highly unbalanced. Second, it allows us to ask for the first time how the predictability of a cascade varies over the range of its growth from small to large. Finally, it more closely approximates the real tasks that need to be solved in applications for managing viral content, where many evolving cascades are being monitored, and the question is which are likely to grow significantly as time moves forward.

For studying cascade growth prediction, it is important to work with a system in which the sharing and resharing of information is widespread, the complete trajectories of many cascades—both large and small—are observable, and the same piece of content shared separately by many people, so that we can begin to control for variation in content. For this purpose, we use a month of complete photo-resharing data from Facebook, which provides a rich ecosystem of shared content exhibiting all of these properties.

In this setting, we focus on several categories of questions:

(i) How high an accuracy can we achieve for cascade growth prediction? If we cannot improve on baseline guessing, then this would be evidence for the inherent unpredictably of cascades. But if we can significantly improve on this baseline, then there is a basis for non-trivial prediction. In the latter case, it also becomes important to understand the features that make prediction possible.

(ii) Is growth prediction more tractable on small cascades or large ones? In other words, does the future behavior of a cascade become more or less predictable as the cascade unfolds?

(iii) Beyond just the growth of a cascade, can we predict its "shape" — that is, its network structure?

Summary of results. Given the challenges in predicting cascades, we find surprisingly strong performance for the growth prediction problem. Moreover, the performance is robust in the sense that multiple distinct classes of features, including those based on time, graph structure, and properties of the individuals resharing, can achieve accuracies well above the baseline. Cascades whose initial reshares come quickly are more likely to grow significantly; and from a structural point of view, breadth rather than depth in the resharing tree is a better predictor of significant growth.

We investigate the performance of growth prediction as a function of the size of the cascade so far — when we want to predict the growth of a cascade of size k, how does our accuracy depend on k? It is not a priori clear whether accuracy should increase or decrease as a function of k, since for any value of k the challenge is to determine what the cascade will do in the future. Seeing more of the cascade (larger k) does not make the problem easier, as it also involves predicting "farther" into the future (i.e., whether the cascade will reach size at least $2k$). We find that accuracy increases

with k, so that it is possible to achieve better performance on large cascades than small ones. The features that are most significant for prediction change with k as well, with properties of the content and the original author becoming less important, and temporal features remaining relatively stable.

We also consider a related question: how much of a cascade do we need to see in order to obtain good performance? Specifically, suppose we want to predict the growth of a cascade of size at least R, but we are only able to see the first $k < R$ nodes in the cascade. How does prediction performance depend on k, and in particular, is there a "sweet spot" where a relatively small value of k gives most of the performance benefits? We find in fact that there is no sweet spot: performance essentially climbs linearly in k, all the way up to $k = R$. Perhaps surprisingly, more information about the cascade continues to be useful even up to the full snapshot of size R.

In addition to growth, we also study how well we can predict the eventual "shape" of the cascade, using metrics for evaluating tree structures as a numerical measure of the shape. We obtain performance significantly above baseline for this task as well; and perhaps surprisingly, multiple classes of features including temporal ones perform well for this task, despite the fact that the quantity being predicted is a purely structural one.

One of the compelling arguments that originally brought the issue of inherent unpredictability onto the research agenda was a striking experiment by Salganik, Dodds, and Watts, in which they showed that the same piece of content could achieve very different levels of popularity in separate independent settings [28]. Given the richness of our data, we can study a version of this issue here in which we can control for the content being shared by analyzing many cascades all arising from the sharing of the same photo. As in the experiment of Salganik et al., we find that independent resharings of the same photo can generate cascades of very different sizes. But we also show that this observation can be compatible with prediction: after observing small initial portions of these distinct cascades for the same photo, we are able to predict with strong performance which of the cascades will end up being the largest. In other words, our data shows wide variation in cascades for the same content, but also predictability despite this variation.

Overall, our goal is to set up a framework in which prediction questions for cascades can be carefully analyzed, and our results indicate that there is in fact a rich set of questions here, pointing to important distinctions between different types of features characterizing cascades, and between the essential properties of large and small cascades.

2. RELATED WORK

Many papers have analyzed and cataloged properties of empirically observed information cascades, while others have considered theoretical models of cascade formation in networks. Most relevant to our work are those which focus on predicting the future popularity of a given piece of content. These studies have proposed rich sets of features for prediction, which we discuss later in Section 3.2.

Much prior work aims to predict the *volume* of *aggregate* activity — the total number of up-votes on Digg stories [29], total hourly volume of news phrases [34], or total daily hashtag use [23]. At the other end of the spectrum, research has focused on *individual* user-level prediction tasks: whether a user will retweet a specific tweet [26] or share a specific URL [10]. Rather than attempt to predict aggregate popularity or individual behavior in the next time step, we instead look at whether an information cascade grows over the median size (or doubles in size, as we later show).

Research on communities defined by user interests [3] or hashtag content [27] has also looked at a notion of growth, predicting

whether a group will increase in size by a given amount. Nevertheless, these focused on groups of already non-trivial size, and their growth predicted without an explicit internal cascade topology, and without tracking predictability over different size classes.

Several papers focus on predictions after having observed a cascade for a given fixed time frame [19, 23, 30]. In contrast, rather than studying specific time slices, we continuously observe the cascade over its entire lifetime and attempt to understand how predictive performance varies as the cascade develops. Moreover, our methodology does not penalize slowly but persistently growing cascades. Thus, we predict the size and the structure after having observed a certain number of initial reshares.

Many studies consider the cascade prediction task as a regression problem [6, 19, 29, 30] or a binary classification problem with large bucket sizes [16, 17, 19]. The danger with these approaches is that they are biased towards studying extremely large but also extremely rare cascades, bypassing the whole issue about the general predictability of cascades. For example, research has specifically focused on content and users that create extremely large cascades, such as popular hashtags [15, 33] and very popular users [9, 14], which has led to criticism that cascades may only be predictable after they have already grown large [31]. While it is useful to understand the dynamics of extremely popular content, such content is also very rare. Thus, we rather seek to understand predictability along cascade's entire lifetime. We consider cascades that have as few as five reshares, and introduce a classification task which is not skewed towards very large cascades.

3. PREDICTING CASCADE GROWTH

To examine the cascade growth prediction problem, we first define and motivate our experimental setup and the feature sets used, then report our prediction results with respect to different k.

3.1 Experimental setup

Mechanics of information passing on Facebook. We focus on content consisting of posts the author has designated as public, meaning that anyone on Facebook is eligible to view it, and we further restrict our attention to content in the form of photos, which comprise the majority of reshare cascades on Facebook [9]. Such posts are then distributed by Facebook's News Feed, typically at first to users who are either friends of the poster or who subscribe to their content, e.g. as followers. Each post is accompanied by a "share" link that allows friends and followers to "reshare" the post with her own friends and followers, thus expanding the set of users exposed to the content. This explicit sharing mechanism creates information cascades, starting with the root node (user or page) that originally created the content, and consisting of all subsequent reshares of that content.

Figure 1 illustrates the process with an example: a node v_0 posts a public photo, seen by v_0's friends and followers in their News Feeds. Friends v_1 and v_3 then share the photo with their own friends. This way the photo propagates over the edges of the Facebook network and creates an information cascade. We represent the cascade graph as $\hat{G}$, and the induced subgraph of all photo sharers, including all friendship or follow links between them as G'. Notice that some users (ex. v_5) are exposed via multiple sources (v_0, v_1, v_3, v_4).

An important issue for our understanding of reshare cascades is the following distinction: content can be produced by *users* — individual Facebook accounts whose primary audience consists of friends and any subscribers the individual has — and it can also be produced by *pages*, which correspond to the Facebook accounts of
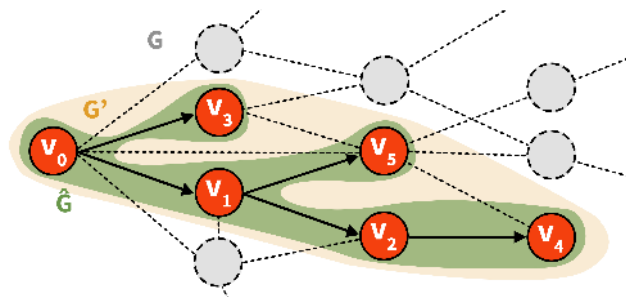

Figure 1: An information cascade represented by solid edges on a graph G, starting at v_0 ($\hat{G}$). Dashed lines indicate friendship edges; the edges between resharers make up the friend subgraph G'.
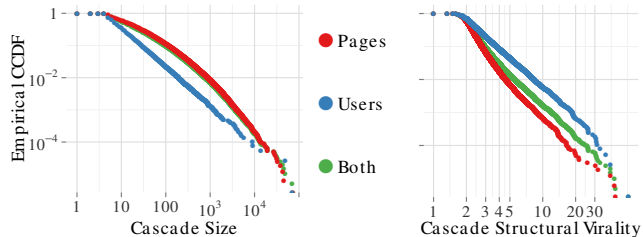

Figure 2: The complementary cumulative distribution (CCDF) of cascade size (left) and structural virality measured by using the Wiener index (right).

companies, brands, celebrities, and other highly visible public entities. In the common parlance around cascades, reshared content originally produced by a user is often informally viewed as more "organic," developing a following in a more bottom-up way. In contrast, reshared content from pages is seen as more top-down, and generally broadcast via News Feed to a larger set of initial followers. A natural question, and a theme that will run through several analyses in the paper, is to understand if these distinctions carry over to the properties we study here: do user-initiated cascades differ in their predictability and their underlying structure from page-initiated cascades?

Dataset description. We sampled our anonymized dataset from photos uploaded to Facebook in June 2013 and observed any reshares occurring within 28 days of initial upload. The dataset only includes photos posted publicly (viewable by anyone), and not deleted during the observation period. Further, we exclude photos with fewer than five reshares as is required by the prediction tasks described below. We constructed diffusion trees first by taking the explicit cascade, e.g. C clicking "share" on B's "share" of A's photo forms the cascade $A \rightarrow B \rightarrow C$. However, it is possible that user C clicked on user B's share, and then directly reshared from A. Since we want to know how the information actually flowed in the network, we reconstruct the path $A \rightarrow B \rightarrow C$ based on click, impression, and friend/follower data [9].

Figure 2 begins to show how photos uploaded by pages generate cascades that differ from those uploaded by users. In our dataset, 81% of cascades are initiated by pages. Figure 2 shows the cascade size distribution for pages, users, and the two combined. Page cascades are typically larger than user cascades, e.g., 11% of page cascades reach at least 100 reshares, while only 2% of user cascades do, though both follow heavy tailed distributions. Fitting power-law curves to their tails, we observe power-law exponents of α equal to 2.2, 2.1, and 2.1 for user, page, and both, respectively ($x_{\min} = 10, 2000, 2000$).

In addition to cascade growth, we quantify the shape of a cascade using the Wiener index, defined as the average distance between all

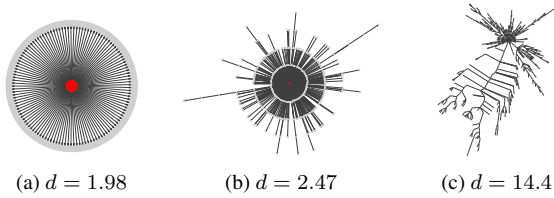| (a) $d = 1.98$ | (b) $d = 2.47$ | (c) $d = 14.4$ |

Figure 3: Cascades with a low Wiener index d resemble star graphs, while those with a high index appear more viral (the root is red).

pairs of nodes in a cascade. Recent work has proposed the Wiener index as a measure of the structural virality of a cascade [2]. Figure 3 shows examples of cascades with varying Wiener index values. Intuitively, a cascade with low structural virality has most of its distribution following from a small number of hub nodes, while a cascade with high virality will have many long paths. Figure 2 shows the distribution of cascade virality (as measured by Wiener index) in our dataset, which, as we saw with cascade size, follows a heavy-tailed distribution. While user cascades are typically smaller than page cascades in our dataset, they tend to have greater structural virality, supporting the intuition that the structure of user-initiated cascades is richer and deeper than that of page-initiated cascades.

Defining the cascade growth prediction problem. Our aim in this paper is to study how well cascades can be predicted. Moreover, we are interested in understanding how various aspects of the prediction task affect the predictive performance.

There are several formulations of the task. If we were to define the task as a regression problem, predictions may be skewed towards large cascades, as cascade size follows a heavy-tailed distribution (Figure 2(right)). Similarly, if we define it as a classification problem of predicting whether a cascade reaches a specific size, we may end up with unbalanced classes, and an overrepresentation of large cascades. Also, if we simply observed a small initial portion of a cascade, and predict its future size, the problem is pathological as almost all cascades are small. And, if we only varied the initial period of observation, the task of predicting whether a cascade reaches a certain size gets easier as we observe more of it.

To remedy these issues, we define a classification task that does not suffer from these deficiencies. We consider a binary classification problem where we observe the first k reshares of a cascade and predict whether the eventual size of a cascade reaches the median size of all the cascades with at least k reshares, $f(k)$. This allows us to study how cascade predictibility varies with k. As exactly half the cascades reach a size greater than the median by definition, random guessing achieves accuracy of 50%.

Interestingly, the question of whether the cascade will reach $f(k)$ is equivalent to that of whether a cascade will double in size. This follows directly from the fact that cascade size distribution follows a power-law with exponent $\alpha \approx 2$. Consider a power-law distribution on the interval (x_{min}, ∞) with a power-law exponent $\alpha \approx 2$. Then the median $f(x)$ of this distribution is $2 \cdot x_{min}$, as demonstrated by the following calculation:

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

As we examine cascades of size greater than $k = x_{min}$, the median size of these cascades is thus $2 \cdot k$ from this derivation. In each of our prediction tasks, we observe that this is indeed true.

Methods used for learning. Our general methodology for the cascade prediction problem will be to represent a cascade by a set of features and then use machine learning classifiers to predict its future size. We used a variety of learning methods, including linear regression, naive Bayes, SVM, decision trees and random forests. However, we primarily report performance of the logistic regression classifier for ease of comparison. In many cases, the performance of most classifiers was similar, although non-linear classifiers such as random forests usually performed slightly better than linear classifiers such as logistic regression. In all cases, we performed 10-fold cross validation and report the classification accuracy, F1 score, and area under the ROC curve (AUC).

3.2 Factors driving cascade growth

We proceed by describing factors that contribute to the growth and spreading of cascades. We group these factors into five classes: properties of the content that is spreading, features of the original poster, features of the resharer, structural features of the cascade, and temporal characteristics of the cascade. Table 1 contains a detailed list of features.

Content features. The first natural factor contributing to the ability of the cascade to spread is the content itself [7]. On Twitter, tweet content and in particular, hashtags, are used to generate content features [23, 30], and identify topics affecting retweet likelihood [26]. LDA topic models have also been incorporated into these prediction tasks [16], and human raters employed to infer the interestingness of content [5, 26]. In our work, we relied on a linear SVM model, trained using image GIST descriptors and color histogram features, to assign likelihood scores of a photo being a closeup shot, taken indoors or outdoors, synthetically generated (e.g., screenshots or pure text vs. photographs), or contained food, a landmark, person, nature, water, or overlaid text (e.g., a meme). We also analyzed words in the caption accompanying an image for positive sentiment, negative sentiment, and sociality [17, 25].

Nevertheless, while content features affected the performance of structural and temporal features, we find that they are weak predictors of how widely disseminated a piece of content would become.

Original poster/resharer features. Some prior work focused on features of the root note in a cascade to predicting the cascade's evolution, finding that content from highly-connected individuals reaches larger audiences, and thus spreads further. Users with large follower counts on Twitter generated the largest retweet cascades [5]. Separately, features of an author of a tweet were shown to be more important than features of the tweet itself [26]. In many Twitter studies predicting cascade size or popularity, a user's number of followers ranks among the top, if not the most, important predictor of popularity [5, 23].

Other features of the root node have also been studied, such as the number of prior retweets of a user's posts [5, 16], and how many Twitter lists a user was included in [26]. The number of @-mentions of a Twitter user was used to predict whether, and how soon a tweet would be retweeted, how many users would directly retweet, and the depth a cascade would reach [33]. Still, [8] found that various measures of a user's popularity are not very correlated with his or her influence.

We capture the intuition behind these factors by defining demographic as well as network features of the original poster as well as the features of the users who reshared the content so far. We use Facebook's distinction of users (individuals) and pages (entities representing an interest) to further distinguish different origin types, in addition to the influence features mentioned above.

Structural features of the cascade. Networks provide the substrate through which information spreads, and thus their structure influences the path and reach of the cascade. As illustrated in Fig-
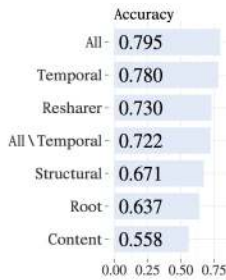
Figure 4: Using logistic regression, we are able to predict with near 80% accuracy whether the size of a cascade will reach the median (10) after observing the first $k =$ 5 reshares.
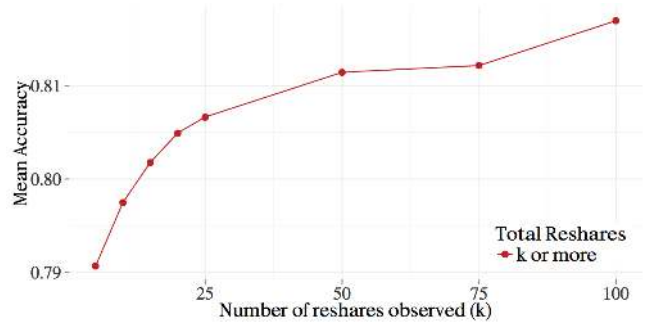

Figure 5: If we observe the first k reshares of a cascade, and want to predict whether the cascade will double in size, our prediction improves as we observe more of it.

ure 1, we generate features from both the graph of the first k reshares ($\hat{G}$), as well as the induced friend subgraph of the first k resharers (G'). Whereas the reshare graph $\hat{G}$ describes the actual spread of a cascade, the friend subgraph G' provides information about the social ties between these initial resharers. The social graph G allows us to compute the potential reach of these reshares.

Previous work considered the network structure of the underlying graph in inferring the virality of content [32], with highly viral items spreading across communities. We use the density of the initial reshare cascade ($subgraph'_k$) and the proximity to the root node ($orig\_connections_k$, did_leave) as proxies for whether an item is spreading primarily within a community or across many. One can also look outside the network between resharers, and count the number of users reachable via all friendship and follow edges of the first k users ($border\_nodes_k$). This relates to total number of exposed users, and has been demonstrated to be an important feature in predicting Twitter hashtag popularity [23].

As we can trace information flow on Facebook exactly, we need not worry about independent entry points influencing a cascade [6, 24]; external influence instead allows us to investigate multiple independent cascades arising from the same content (see Section 5.1).

Temporal features. Properties related to the "speed" of the cascade (e.g., $time_k$) were shown to be the most important features in predicting thread length on Facebook [4], and are a primary mechanism in predicting online content popularity [29]. Moreover, as the speed of diffusion changes over time, this may have a strong effect on the ability of the cascade to continue spreading through the network [33].

We characterize a number of temporal properties of cascade diffusion (see Table 1). In particular, we measure the change in the speed of reshares ($time''_{1..k}$), compare the differences between the speed in the first and second half of the measurement period ($time'_{1..k/2}$, $time'_{k/2..k}$), and qunatify the number of users who were exposed to the cascade per time unit ($views'_{1..k-1,\ k}$).

3.3 Predicting cascade growth

To illustrate the general performance of the features described in the previous section we consider a simple prediction task, where we observe the first 5 reshares of the cascade and want to predict whether it will reach the median cascade size (or equivalently, whether it will double and be reshared at least 10 times). For the experiment we use a set of $N_c = 150,572$ photos, where each photo was shared at least 5 times. The total number of reshares of these photos was $N_r = 9,233,300$.

Figure 4 shows logistic regression performance using all features from Table 1. For this task, random guessing would obtain a performance of 0.5, while our method achieves surprisingly strong per-

formance: classification accuracy of 0.795 and AUC of 0.877. If we relax the task and instead of predicting above vs. below median size, we predict top vs. bottom quartile (top 25% vs. bottom 25%) the accuracy rises even further to 0.926, and the AUC to 0.976.

Overall, while each feature set is individually significantly better than predicting at random, it is the set of temporal features that outperforms all other individual feature sets, obtaining performance scores within 0.025 of those obtained when using all features. To understand if we could do well without temporal features, we trained a classifier which excluded them and were still able to obtain reasonable performance even without these features. This is especially useful when one knows through *whom* information was passed, but not *when* it was passed. The lack of reliance on any individual set of features demonstrates that the predictions are robust.

Studied individually, we also find that temporal features generally performed best, followed by structural features. The reshare rate in the second half ($time'_{k/2..k}$) was most predictive, attaining accuracy of 0.73. This was followed by the rate of user views of the original photo, $views'_{0,k}$, and the time elapsed between the original post and fifth reshare, $time_5$ (both 0.72). In fact, $time_{k+1}$ is always more accurate than $time_k$. The most accurate structural features were did_leave and $outdeg(v_0)$ (both 0.65). We examine individual feature importance in more detail later.

3.4 Predictability and the observation window of size k

It is also natural to ask whether cascades get more or less predictable as we observe more of the initial growth of a cascade. One may think that observing more of the cascade may allow us to extrapolate its future growth better; on the other hand, additional observed reshares may also introduce noise and uncertainty in the future growth of the cascade. Note that the task does not get easier as we observe more of the cascade, as we are predicting whether the cascade will reach size $2k$ (or equivalently, the median) given that we have seen k reshares so far.

Figure 5 shows that the predictive performance of whether a cascade doubles in size increases as a function of the number of observed reshares k. In other words, it is easier to predict whether a cascade that has reached 25 reshares will get another 25, than to predict whether a cascade that has reached 5 reshares will obtain an additional 5. Thus, the prediction accuracy for larger cascades is above the already high accuracy for smaller values of k. The change in the F1 score and AUC also follow a very similar trend.

Overall, these results demonstrate that observing more of the cascade, while also predicting "farther" into the future, is easier than observing a cascade early in its life and predicting what it will

Content Features

$score_{food/nature/...}$	The probability of the photo having a specific feature (food, overlaid text, landmark, nature, etc.)
is_en	Whether the photo was posted by an English-speaking user or page
$has\_caption$	Whether the photo was posted with a caption
$liwc_{pos/neg/soc}$	Proportion of words in the caption that expressed positive or negative emotion, or sociality, if English

Root (Original Poster) Features

$views_{0,\,k}$	Number of users who saw the original photo until the kth reshare was posted
$orig\_is\_page$	Whether the original poster is a page
$outdeg(v_0)$	Friend, subscriber or fan count of the original poster
age_0	Age of the original poster, if a user
$gender_0$	Gender of the original poster, if a user
fb_age_0	Time since the original poster registered on Facebook, if a user
$activity_0$	Average number of days the original poster was active in the past month, if a user

Resharer Features

$views_{1..k-1,\,k}$	Number of users who saw the first $k-1$ reshares until the kth reshare was posted
$pages_k$	Number of pages responsible for the first k reshares, including the root, or $\sum_{i=0}^{k} \mathbb{1}\{v_i \text{ is a page}\}$
$friends_k^{avg/90p}$	Average or 90th percentile friend count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{friends}(v_i)\mathbb{1}\{v_i \text{ is a user}\}$
$fans_k^{avg/90p}$	Average or 90th percentile fan count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg(v_i)\mathbb{1}\{v_i \text{ is a page}\}$
$subscribers_k^{avg/90p}$	Average or 90th percentile subscriber count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{subscriber}(v_i)\mathbb{1}\{v_i \text{ is a user}\}$
$fb\_ages_k^{avg/90p}$	Average or 90th percentile time since the first k resharers registered on Facebook, or $\frac{1}{k}\sum_{i=1}^{k} fb\_age_i$
$activities_k^{avg/90p}$	Average number of days the first k resharers were active in July, or $\frac{1}{k}\sum_{i=1}^{k} activity_i$
$ages_k^{avg/90p}$	Average age of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} age_i$
$female_k$	Number of female users among the first k resharers, or $\sum_{i=1}^{k} \mathbb{1}\{gender_i \text{ is female}\}$

Structural Features

$outdeg(v_i)$	Connection count (sum of friend, subscriber and fan counts) of the ith resharer (or out-degree of v_i on $G = (V, E)$)
$outdeg(v_i')$	Out-degree of the ith reshare on the induced subgraph $G' = (V', E')$ of the first k resharers and the root
$outdeg(\hat{v}_i)$	Out-degree of the ith reshare on the reshare graph $\hat{G} = (\hat{V}, \hat{E})$ of the first k reshares
$orig\_connections_k$	Number of first k resharers who are friends with, or fans of the root, or $|\{v_i \mid (v_0, v_i) \in E,\ 1 \leq i \leq k\}|$
$border\_nodes_k$	Total number of users or pages reachable from the first k resharers and the root, or $|\{v_i \mid (v_i, v_j) \in E,\ 0 \leq i, j \leq k\}|$
$border\_edges_k$	Total number of first-degree connections of the first k resharers and the root, or $|\{(v_i, v_j) \mid (v_i, v_j) \in E,\ 0 \leq i, j \leq k\}|$
$subgraph_k'$	Number of edges on the induced subgraph of the first k resharers and the root, or $|\{(v_i, v_j) \mid (v_i, v_j) \in E',\ 0 \leq i, j \leq k\}|$
$depth_k'$	Change in tree depth of the first k reshares, or $\min_\beta \sum_{i=1}^{k}(depth_i - \beta i)^2$
$depths_k^{avg/90p}$	Average or 90th percentile tree depth of the first k reshares, or $\frac{1}{k}\sum_{i=1}^{k} depth_i$
did_leave	Whether any of the first k reshares are not first-degree connections of the root

Temporal Features

$time_i$	Time elapsed between the original post and the ith reshare
$time_{1..k/2}'$	Average time between reshares, for the first $k/2$ reshares, or $\frac{1}{k/2-1}\sum_{i=1}^{k/2-1}(time_{i+1} - time_i)$
$time_{k/2..k}'$	Average time between reshares, for the last $k/2$ reshares, or $\frac{1}{k/2-1}\sum_{i=k/2}^{k-1}(time_{i+1} - time_i)$
$time_{1..k}''$	Change in the time between reshares of the first k reshares, or $\min_\beta \sum_{i=1}^{k-1}(time_{i+1} - time_i) - \beta i)^2$
$views_{0,k}'$	Number of users who saw the original photo, until the kth reshare was posted, per unit time, or $\frac{views_{0,\,k}}{time_k}$
$views_{1..k-1,\,k}'$	Number of users who saw the first $k-1$ reshares, until the kth reshare was posted, per unit time, or $\frac{views_{1..k-1,k}}{time_k}$

Table 1: List of features used for learning. We compute these features given the cascade until the kth reshare.

do next (i.e., $k = 5$ vs. $k = 25$).

Fixing the minimum cascade size R. In the previous version of the task, cascades are required only to have at least k reshares. Thus, the set of cascades changes with k. Here, we examine a variation of this task, where we compose a dataset of cascades that have at least R reshares. We observe the first k ($k \leq R$) reshares of the cascade and aim to predict whether the cascade will grow over the median size (over all cascades of size $\geq R$). As we increase k, the task gets easier as we observe more of the cascade and the predicted quantity does not change.

With the task, we find that performance increases linearly with

k up to R, or that there is no "sweet spot" or region of diminishing returns ($p < 0.05$ using a Harvey-Collier test). For example, the top-most line in Figure 6 shows that when each observed cascade has obtained 100 or more reshares, performance increases linearly as more of the cascade is observed. This demonstrates that more information is always better: the greater the number of observed reshares, the better the prediction.

However, Figure 6 also shows that larger cascades are less predictable than smaller cascades. For example, predicting whether cascades with 1,000 to 2,000 reshares grow large is significantly more difficult than predicting cascades of 100 to 200 reshares. This
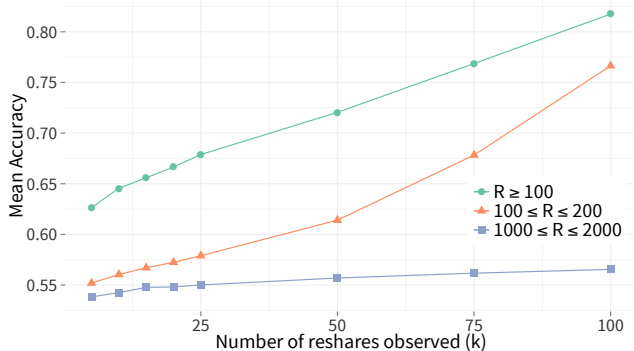
Figure 6: Knowing that a cascade obtains at least R reshares, prediction performance increases linearly with k, $k \leq R$. However, differentiating among cascades with large R also becomes more difficult.

shows that once one knows that a cascade will grow to be large, knowing the characteristics of the very beginning of its spread is less useful for prediction.

3.5 Changes in feature importance

We now examine how feature importance changes as more and more of the cascade is observed. In this experiment, we compute the value of the feature after observing first k reshares and measure the correlation coefficient of the feature value with the log-transformed number of reshares (or cascade size).

Figure 7 shows the results for the five feature types. We summarize the results by the following observations:

- *Correlations of averages increase with the number of observations.* As we obtain more examples, naturally averages get less noisy, and more predictive (e.g., $ages^{avg}$ and $friends^{avg}$).

- *The original post gets less important with increasing k.* After observing 100 reshares, it becomes less important that the original post was made by a page ($orig\_is\_page$), or that the original poster had many connections to other users ($outdeg(v_0)$).

- *Similarly, the actual content being reshared gets less important with increasing k.* Almost all content features tend to zero as k increases, except for $has\_caption$ and is_en. This can be explained by the fact that cascades of photos with captions have a unimodal distribution, and cascades started by English speakers have a bimodal distribution. Thus, these features become correlated in opposite directions.

- *Successful cascades get many views in a short amount of time, and achieve high conversion rates.* The number of users who have viewed reshares of a cascade is more negatively correlated with increasing k ($views_{1..k-1,k}$), suggesting that requiring "fewer tries" to achieve a given number of reshares is a positive indicator of its future success. On the other hand, while requiring fewer views is good, rapid exposure, or reaching many users within a short amount of time is also a positive predictor ($views'_{1..k-1,k}$).

- *Structural connectedness is important, but gets less important over time.* Nevertheless, reshare depth remains highly correlated: the deeper a cascade goes, the more likely it is to be long-lasting, as even users "far away" from the original poster still find the content interesting.
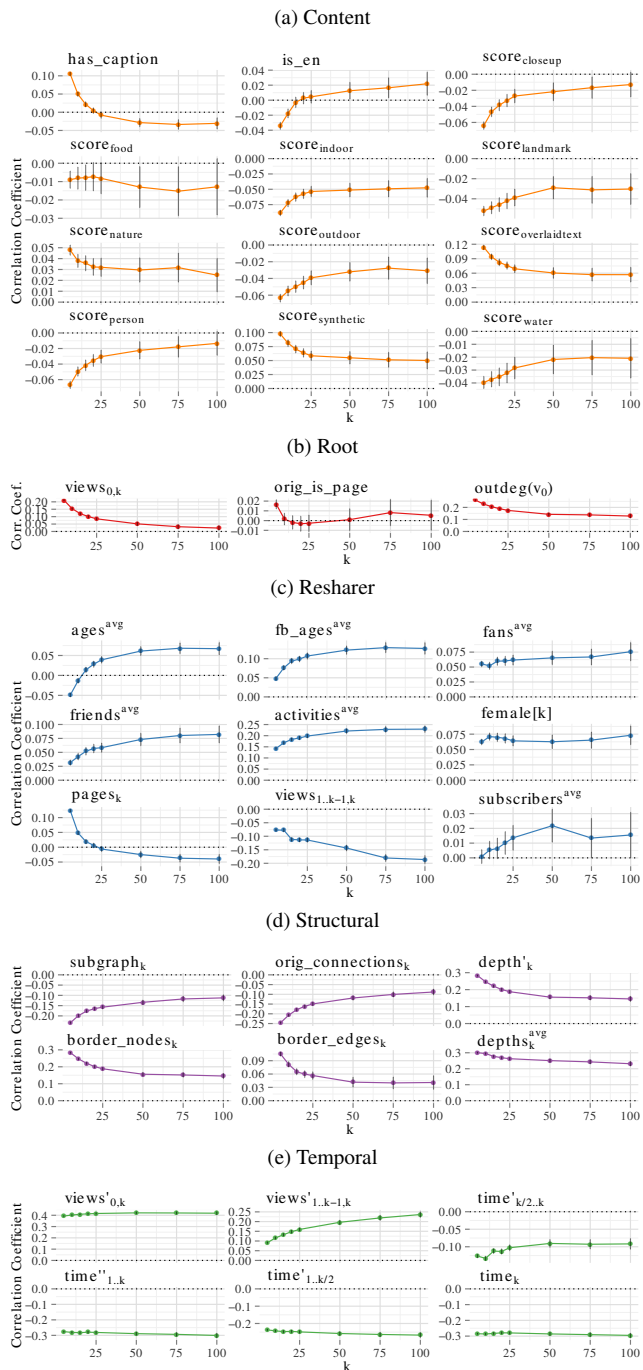

Figure 7: The importance of each feature varies as we observe more of a cascade, as shown by the change in correlation coefficients.

- *The importance of timing features remains relatively stable.* While highly correlated, timing features remain remarkably stable in importance as k increases.

We note individual features' logistic regression coefficients empirically follow similar shapes, but have the downside of having interactions with one another. Using either the slope of the best-fit line of the cascade size against the normalized feature value, or individual feature performance also reveals similar trends. Further LIWC text content features (positive, negative, and social categories) consistently performed poorly, attaining performance no
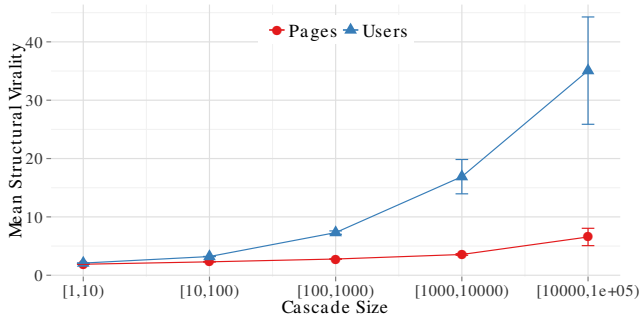
Figure 8: The mean structural virality (Wiener index) increases with cascade size, but is significantly higher for user cascades.

better than chance, with accuracy between 0.49 and 0.52.

4. PREDICTING CASCADE STRUCTURE

Similar to predicting cascade size we can also attempt to predict the *structure* of the cascade. We now turn to examining how structural features of the cascade determine its evolution and spread.

4.1 User-started and page-started cascades

Earlier we discussed the notion of *structural virality* as a measure of how much the structure of a cascade is dominated by a few hub nodes, and we saw that user-initiated cascades have significantly higher structural virality than page-initiated cascades, reflecting their richer graph structure. It is natural to ask how these distinctions vary with the size of the cascade — are large user-initiated cascades more similar to page-initiated ones, e.g. are they driven by popular hub nodes?

Figure 8 shows that the opposite is the case — user and page-initiated cascades remain structurally distinct, with this distinction even increasing with cascade size. Moreover, this difference continues to hold even when controlling for the number of first-degree reshares (directly from the root), suggesting a certain robustness to their richer structure. Because of these structural differences, we handle user and page cascades separately in the analyses that follow.

These distinctions may also help explain a large difference in the predictability of user-initiated vs. page-initiated cascades. We observe that for page cascades accuracy exceeds 80%, while that for user cascades is slightly under 70%. (These results also hold for the F1 score and AUC, with a difference of about 0.1.) The fact that much more of the structure of a page-initiated cascade is typically carried by a small number of hub nodes may suggest why the prediction task is more tractable in this case.

4.2 The initial structure of a cascade influences its eventual size

To understand how structure bears on the future growth of the cascade, we examine how the configuration of the first three reshares (and the root) correlates with the cascade size. In particular, we measure the proportion of cascades starting from each configuration that reach the median size. We do this separately for two different initial poster types: a user, and a page. We discard "celebrity" users who may large followings like the most popular pages. Figure 9a shows that as the initial cascade structure becomes shallower, the proportion of cascades that double in size increases. To examine why this would be the case, we also examined the time needed for the 3rd reshare to happen (Figure 9c). For pages, shallower cascades tend to happen more rapidly, consistent with being
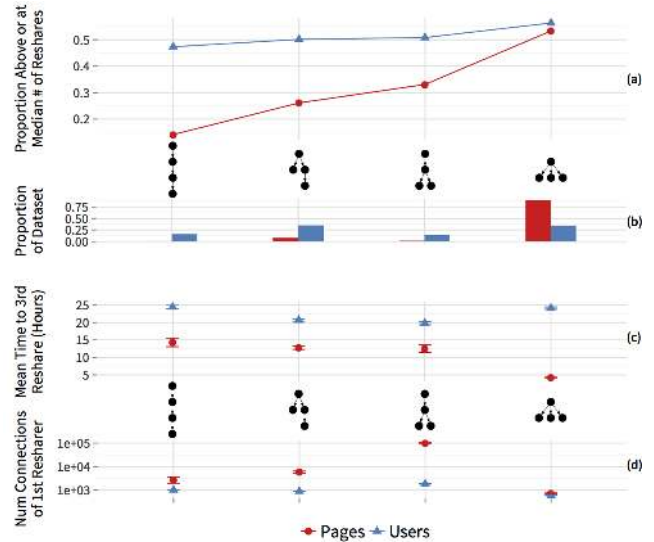

Figure 9: Shallow initial cascade structures are indicative of larger cascades. In contrast to page-started cascades, where the mean time to the 3rd reshare decreases with decreasing depth of the initial cascade, shallow cascades take a much longer time to form for user-started cascades. For these, the connections of the 1st resharer also significantly impacts the time to the 3rd resharer, especially when it receives two reshares before the original receives a second.

initiated by a popular page and achieving a large number of reshares directly from its fans. Interestingly, the configuration having the second and third reshares stemming from the first reshare correspond to having a first resharer with many connections, and indicating that the initial poster is less popular, be it a page or user (Figure 9d).

Curiously, for user-started cascades, the star configuration tends to grow into the largest cascades, but is also the slowest. It also tends to correspond to the first resharer having a low degree, both for page and user roots. One might speculate that this pattern is indicative of the item's appeal to less well-connected users, who also happen to be more likely to reshare. In fact, a median resharer has 35 fewer friends than someone who is active on the site nearly every day. Thus, an item's appeal, rather than the initial network structure, may drive the eventual cascade size in the long run.

4.3 Predicting cascade structure

The observations above naturally lead to the question of whether it is possible to predict future cascade structure. In particular, we aim to distinguish cascades that spread like a virus in a shallow forest fire-like pattern (Figure 3a) and cascades which spread in long, narrow string-like pattern (Figure 3c). As discussed earlier, this difference is related to the structural virality of a cascade and is quantified by the Wiener index. Here, we observe $k = 5$ reshares of a cascade and aim to predict whether the final cascade will have a Wiener index above or below the median. We obtain accuracy of 0.725 (F1 = 0.715, AUC = 0.796), while random guessing would, by construction, achieve accuracy of 0.5.

Temporal and structural features are most predictive of structure. For this task we expect structural features to be most important, while we expect temporal features not to be indicative of the cascade structure. However, when we train the model on individual classes of features we surprisingly find that both temporal and structural features are almost equally useful in predict-
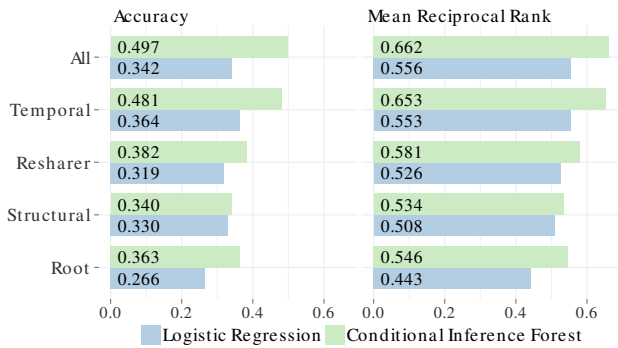
Figure 10: In predicting the largest cascade in clusters of 10 or more cascades of identical photos, we perform significantly above the baseline of 0.1.
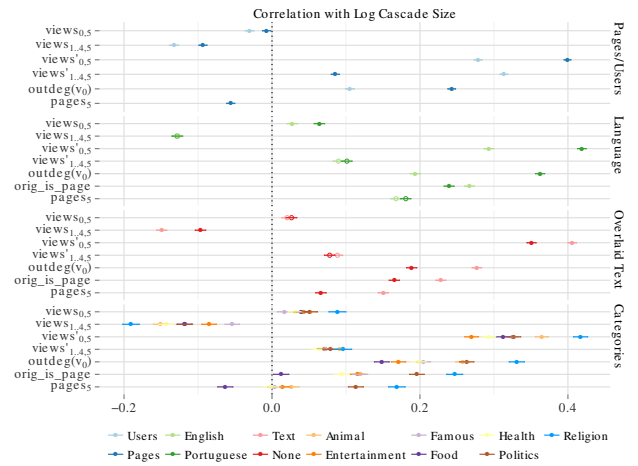

Figure 11: The initial exposure of the uploaded photo and initial reshares serve to differentiate datasets from one another, as can be seen by comparing the correlation coefficients of each feature with the log cascade size. Solid circles indicate significance at $p < 10^{-3}$, and lines through each circle indicate the 95% confidence interval.

ing cascade structure: 0.622 vs. 0.620. Nevertheless, structural features remain individually more accurate (≈ 0.58) and highly correlated ($0.161 \leq |r| \leq 0.255$) with the Wiener index. Individually, one temporal feature, $views'_{1..k-1,k}$, is slightly more accurate (0.602) compared to the best-performing structural feature, $outdeg(\hat{v}_0)$ (0.600), but is significantly less correlated (0.041 vs. -0.255). The two classes of features nicely complement each other, since when combined, accuracy increases to 0.72.

Cascade structure also becomes more predictable with increasing k. Like for cascade growth prediction, our prediction performance improves as we observe more of the cascade, with accuracy linearly increasing from 0.724 when k is 5 to 0.808 when k is 100. A linear relation also exists in the alternate task where we set the minimum cascade size R to be 100, varying k between 5 and 100.

Changes in feature importance. As we increase k, we find that the structural features become highly correlated with the Wiener index, suggesting that the initial shape of a cascade is a good indicator of its final structure. Rapidly growing cascades also result in final structures that are shallower—temporal features become more strongly correlated with the Wiener index as k increases. Unlike with cascade size, views were generally weakly correlated with structure, while content features had a weak, near-constant effect. Nonetheless, some of these features still provided reasonable performance in the prediction task.

User vs. page-started cascades. In predicting the shape of a cascade, we find that our overall prediction accuracy for pages is slightly higher (0.724) than for users (0.700). While using only structural features alone results in a higher prediction accuracy for users (0.643) than for pages (0.601), user and content features are significantly more predictive of cascade structure in the case of pages.

To sum up, we find that predicting the shape of a cascade is not as hard as one might fear. Nevertheless, predicting cascade size is still much easier than predicting cascade shape, though classifiers for either achieve non-trivial performance.

5. PREDICTABILITY & CONTENT

5.1 Controlling for cascade content

In our analyses thus far, we examined cascades of uploads of different photos, and tried to account for content differences by including photo and caption features. However, temporal and structural features may still capture some of the difference in content. Thus, we now study how well we can predict cascade size if we control for the content of the photo itself. We consider *identical*

photos uploaded to Facebook by different users and pages, which is not a rare occurrence. We used an image matching algorithm to identify copies of the same image and place their corresponding cascades into clusters (983 clusters, $N_c = 38,073$, $N_r = 12,755,621$). As one might expect, even the same photo uploaded at different times by different users can fare dramatically differently; a cluster typically consists of a few or even a single cascade with a large number of reshares, and many smaller cascades with few reshares. The average Gini coefficient, a measure of inequality, is 0.787 ($\sigma = 0.104$) within clusters. Thus, a natural task is to try to predict the largest cascade within a cluster. For every cluster we select 10 random cascades, placing the accuracy of random guessing at 10%.

As shown in Figure 10, in all cases we significantly outperform the baseline. Using a random forest model, we can identify the most popular cascade nearly half the time (accuracy 0.497); a mean reciprocal rank of 0.662 indicates that this cascade also appears in the top two predicted cascades almost all the time.

In terms of feature importance we notice that best results are obtained using temporal features, followed by resharer, root node, and structural features. Essentially, if one upload of the photo is initially spreading more rapidly than other uploads of the same photo, that cascade is also likely to grow to be the largest. This points to the importance of landing in the right part of the network at the right time, as the same photo tends to have widely and predictably varying outcomes when uploaded multiple times.

5.2 Feature importance in context

Some features may be more or less important for our prediction tasks in different contexts. Figure 11 shows how several features correlate with log-transformed cascade size when conditioned on one of four different variables, including (1) source node type—user vs page, (2) language—English versus Portuguese, the two most common languages of cascade root nodes in our dataset, (3) whether text is overlaid on a photo—a common feature of recent Internet memes, and (4) content category. We determine content category by matching entities in photo captions to Wikipedia articles, and in turn articles to seven higher-level categories: animal, entertainment, politics, religion, famous people (excluding religious and political figures), food, and health.

Figure 11 shows that the initial rate of exposure of the uploaded photo is generally more important for page cascades than for user cascades ($views'_{0,5}$). This is likely due to the higher variance in the distribution of the number of followers for a user versus a page. For page cascades in our sample, the median number of followers is 73,855 with a standard deviation of 675,203, while for users at the root of cascades the median number of friends and subscribers is 1,042 with a standard deviation of 26,482. Though rate of exposure to the original photo is more important for pages, we see that rate of exposure to the initial reshares ($views'_{1..4,5}$) is much more important for user cascades.

The number and rate of views also act to differentiate topical categories, with religion having the highest correlation between views and cascade size. Correlation for the rate of views of the uploaded photo is also higher for those with a Portuguese-speaking root node as opposed to an English one. The feature $outdeg(v_0)$ indicates the ability of the root to broadcast content, and we see this playing an important role for page cascades, Portuguese content, photos with text, and religious photos. This indicates that much of the success of these cascades is related to the root nodes being directly connected to large audiences.

In addition to the analysis of Figure 11, we also examined how the features correlate with the structural virality of the final cascades. (Each of the reported correlation coefficient comparisons that follow are significant at $p < 10^{-3}$ using a Fisher transformation.) Photos relating to food differ significantly from all other categories in that features of the root, such as $outdeg(v_0)$, are less negatively correlated ($>$–0.18 vs. –0.11), and depth features, such as $depth_k^{avg}$, are less positively correlated ($>$0.18 vs. 0.11). This relationship also holds for English compared to Portuguese photos. While users with many friends or followers are more likely to generate cascades of larger size and greater structural virality, pages with many fans create cascades of larger size, although not necessarily greater virality (0.05 vs. –0.01). However, if the initial structure of a cascade is already deep, the final structure of the cascade is likely to have greater structural virality for both user and page-started cascades ($>$0.16). A user-started cascade whose initial reshares are viewed more quickly is also more likely to become viral than that for a page-started cascade (0.23 vs. 0.06).

6. DISCUSSION & CONCLUSION

This paper examines the problem of predicting the growth of cascades over social networks. Although predictive tasks of similar spirit have been considered in the past, we contribute a novel formulation of the problem which does not suffer from skew biases. Our formulation allows us to study predictability throughout the life of a cascade. We examine not only how the predictability changes as more and more of the cascade is observed (it improves), but also how predictable large cascades are if we only observe them initially (larger cascades are more difficult to predict). While some features, e.g., the average connection count of the first k resharers, have increasing predictive ability with increasing k, others weaken in importance, e.g., the connectivity of the root node. We find that the importance of features depends on properties of the original upload as well: the topics present in the caption, the language of the root node, as well as the content of the photo.

Despite the rich set of results we were able to obtain, there are some limitations to this study. Most importantly, the study was conducted entirely with Facebook data and only with photos. Still, one advantage of this is the scale of the medium; hundreds of millions of photos are uploaded to Facebook every day, and photos, more than other content types, tend to dominate reshares. This also gives us high-fidelity traces of how the photo moves within Facebook's
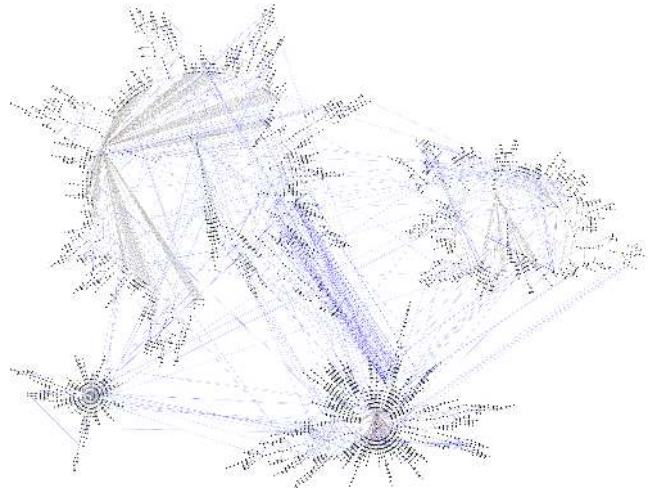

Figure 12: There is considerable overlap in friendship edges (blue) between four independent cascades of the same photo.

ecosystem, which allows us to precisely overlay the spreading cascade over the social network. Moreover, we are able to identify uploads of the same photo and track them individually. This eliminates the concern of shares being driven by an external entity and only appearing to be spreading over the network. Instead, external drivers benefit our study by creating independent 'experiments' where the same photo gets multiple chances to spread, helping us control for the role of content in some of our experiments. Another disadvantage of our setup is that diffusion within Facebook is driven by the mechanics of the site. The distinction between pages and users is specific to Facebook, as are the mechanisms by which users interact with content, e.g., liking and resharing. Despite these limitations, we believe the results give general insights which will be useful in other settings.

The present work only examines each cascade independently from others. Future work should examine interactions between cascades, both between different content competing for the same attention, and between the same content surfacing at different times and in different parts of the network. We found that when the same photo is uploaded at least 10 times, the largest cascade was twice as likely to be among the first 20% of uploads than the last 20%. Similarly, for photos uploaded 20 times, the largest cascade was 2.3 times as likely to be among the first 20% than the last. Figure 12 shows the friendship edges between users participating in different cascades of a single, specific photo. The high connectivity *between* different cascades demonstrates that users are likely being exposed to the same photo via different cascades, which could be a contributing factor in why earlier uploads of the same photo tend to generate larger cascade than later ones. Between-cascade dynamics like this should provide ample opportunities for further research.

Addressing questions like these will lead to a richer understanding of how information spreads online and pave the way towards better management of socially shared content and applications that can identify trending content in its early stages.

7. REFERENCES

[1] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, 2004.

[2] A. Anderson, S. Goel, J. Hofman, and D. Watts. The structural virality of online diffusion. *Under review*.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

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

[5] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proc. WSDM*, 2011.

[6] E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In *Proc. EC*, 2009.

[7] J. Berger and K. L. Milkman. What makes online content viral. *J. Marketing Research*, 49(2):192–205, 2012.

[8] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proc. ICWSM*, 2010.

[9] P. A. Dow, L. A. Adamic, and A. Friggeri. The anatomy of large facebook cascades. In *Proc. ICWSM*, 2013.

[10] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers-predicting information cascades in microblogs. In *Proc. OSM*, 2010.

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

[12] B. Golub and M. O. Jackson. Using selection bias to explain the observed structure of internet diffusions. *Proc. Natl. Acad. Sci.*, 2010.

[13] D. Gruhl, R. V. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. WWW*, 2004.

[14] M. Guerini, J. Staiano, and D. Albanese. Exploring image virality in google plus. *Proc. SocialCom*, 2013.

[15] T.-A. Hoang and E.-P. Lim. Virality and susceptibility in information diffusions. In *Proc. ICWSM*, 2012.

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

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

[18] R. Kumar, M. Mahdian, and M. McGlohon. Dynamics of conversations. In *Proc. KDD*, 2010.

[19] 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.

[20] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 2007.

[21] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *Proc. ICDM*, 2007.

[22] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl. Acad. Sci.*, 2008.

[23] Z. Ma, A. Sun, and G. Cong. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the American Society for Information Science and Technology*, 2013.

[24] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proc. KDD*, 2012.

[25] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: LIWC 2001. 2001.

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

[27] D. M. Romero, C. Tan, and J. Ugander. On the interplay between social and topical structure. In *Proceedings of the Seventh International Conference on Weblogs and Social Media (ICWSM)*, 2013.

[28] M. Salganik, P. Dodds, and D. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 2006.

[29] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 2010.

[30] 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.

[31] D. J. Watts. *Everything is Obvious: How Common Sense Fails Us*. Crown, 2012.

[32] L. Weng, F. Menczer, and Y.-Y. Ahn. Virality prediction and community structure in social networks. *Sci. Rep.*, 2013.

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

[34] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *Proc. ICDM*, 2010.