Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Predictability of Popularity: Gaps between Prediction and Understanding

Benjamin Shulman
Dept. of Computer Science
Cornell University
bgs53@cornell.edu

Amit Sharma
Microsoft Research
New York, NY
amshar@microsoft.com

Dan Cosley
Information Science
Cornell University
drc44@cornell.edu

Abstract

Can we predict the future popularity of a song, movie or tweet? Recent work suggests that although it may be hard to predict an item's popularity when it is first introduced, peeking into its early adopters and properties of their social network makes the problem easier. We test the robustness of such claims by using data from social networks spanning music, books, photos, and URLs. We find a stronger result: not only do predictive models with peeking achieve high accuracy on all datasets, they also generalize well, so much so that models trained on any one dataset perform with comparable accuracy on items from other datasets.

Though practically useful, our models (and those in other work) are intellectually unsatisfying because common formulations of the problem, which involve peeking at the first small-k adopters and predicting whether items end up in the top half of popular items, are both too sensitive to the speed of early adoption and too easy. Most of the predictive power comes from looking at how quickly items reach their first few adopters, while for other features of early adopters and their networks, even the direction of correlation with popularity is not consistent across domains. Problem formulations that examine items that reach k adopters in about the same amount of time reduce the importance of temporal features, but also overall accuracy, highlighting that we understand little about why items become popular while providing a context in which we might build that understanding.

How does a book, song, or a movie become popular? The question of how cultural artifacts spread through social networks has captured the imagination of scholars for decades. Many factors are cited as important for an item to spread *virally* through social networks and become popular: its intrinsic quality (Gladwell 2006a; Simonoff and Sparrow 2000), the characteristics of its initial adopters (Gladwell 2006b), the emotional response it elicits (Berger and Milkman 2012), and so on. Often, explanations are used to justify the popularity of different items after the fact (Berger 2013), making it hard to apply these explanations to new events (Watts 2011).

Online social networks allow us to observe individual-level traces of how items are transferred between people, allowing more precise modeling of the phenomenon. Predicting the future popularity of an item based on attributes of the item and the person who introduced it has emerged as a useful problem, both to understand processes of information diffusion and to inform content creation and feed design on social media platforms. For example, Twitter's managers may want to highlight new tweets that are more likely to become popular, while its users may want to learn from characteristics of popular tweets to imporve their own.

In general, even with detailed information about an item's content or the person sharing it, it is hard to predict which items will become more popular than others (Bakshy et al. 2011; Martin et al. 2016). The problem becomes more tractable when we are allowed to *peek* into the initial spread of an item. The intuition is that early activity data about the speed of adoption, characteristics of people who adopt it and the connections between them might predict the item's fate. This intuition shows encouraging results for both predicting the final popularity of an item (Szabo and Huberman 2010; Pinto, Almeida, and Gonçalves 2013; Zhao et al. 2015) and whether an item will end up in the top 50% of popular items (Cheng et al. 2014; Romero, Tan, and Ugander 2013; Weng, Menczer, and Ahn 2013).

Buoyed by these successes, one might conclude that the availability of rich features about the item and social network of early adopters has helped us understand why items become popular. However, past work studies individual datasets and varying versions of the prediction problem, making it hard to compare results. For instance, studies disagree on the direction of the effect of network structural features on item popularity (Lerman and Hogg 2010; Romero, Tan, and Ugander 2013).

In this paper, we try to unify these observations on popularity prediction through studying different problem formulations and kinds of features over a wide range of online social networks. Using an existing formulation that predicts whether the final popularity of items is above the median based on features of the first five adopters (Cheng et al. 2014), we confirm past work (Szabo and Huberman 2010) showing that features about those adopters and their social network are at best weak predictors of popularity compared to temporal features. For instance, a single temporal heuristic—the average rate of early adoption—is a better predictor than all non-temporal features combined across all four websites. Further, models trained on one dataset and tested on others using temporal features generalize fairly well, while those that use network structural features generalize badly.

In one reading, this is a useful contribution: peeking-based popularity models that include temporal information achieve up to 83% accuracy on Twitter and generalize well across datasets. From a practical standpoint, we encourage content distributors to use temporal features for predicting the future success of items.

Intellectually, however, our finding is not very satisfying. Rather than identifying features that shed light on why items become popular, we mostly see that items that become popular fast are more likely to achieve higher popularity in the end. Rapid adoption may be a signal of quality, interestingness, and eventual popularity—but doesn't tell us why. The effect might also be driven by cumulative advantage (Frank and Cook 2010; Watts 2011): items that receive attention early have more chances to spread through via interfaces that highlight popular or trending items.

An alternative formulation of the problem that reduces the effect of temporal features lets us see just what early adopter and network features tell us. This formulation, called Temporal Matching, compares items that achieve similar levels of popularity in the same amount of time, rather than the more common formulation of looking at the first k adopters regardless of the time it takes to reach k. Controlling for the average rate of an item's adoption turns popularity prediction into a hard problem. Using the same features as before, prediction accuracy across all datasets drops below 65%. Such a decrease in accuracy underscores the importance of choosing problem formulations that highlighting relevant phenomena in popularity evolution. Current models may fare well on certain formulations, but there is still much to learn about how items become popular.

Formulations of the prediction problem

We start by identifying two key dimensions to consider when defining the popularity prediction task: how much peeking into early activity on an item is allowed, and whether the task is a regression or classification. For ease of exposition, we use *item* to denote entities that are consumed in online social networks. *Adoption* refers to an explicit action or endorsement of an item, such as loving a song, favoriting a photo, rating a book highly or retweeting a URL. Finally, we define *popularity* of an item as the number of people who have adopted it.

Predicting apriori versus peeking into early activity

Predicting popularity *a priori* for items such as movies (Simonoff and Sparrow 2000) or songs (Pachet and Sony 2012) has long been considered a hard problem. One of the most successful approaches has been to gauge audiences' interest in an item before it is officially released, such as by measuring the volume of tweets (Asur, Huberman, and others 2010) or search queries (Goel et al. 2010). Such methods can work well for mainstream, popular items for which there might be measurable prior buzz, but are unlikely to be useful for genuinely new items such as tweets or photos uploaded by users.

For such items, popularity prediction is tricky, even when precise data about the content of each tweet and the seed user's social network is known. On Twitter, models with extensive content features such as the type of content, its source and topic, crowdsourced scores of interestingness, and features about the seed user such as indegree and past popularity of tweets are only able to explain less than half of the variance in popularity (Martin et al. 2016). Further, the content features are usually less important than features of the seed user (Bakshy et al. 2011; Martin et al. 2016; Jenders, Kasneci, and Naumann 2013).

In response, scholars have suggested modified versions of the problem where one peeks into early adoption activity for an item. In studies on networks including Facebook (Cheng et al. 2014), Twitter (Lerman and Hogg 2010; Zhao et al. 2015; Tsur and Rappoport 2012; Kupavskii et al. 2013), Weibo (Yu et al. 2015), Digg (Lerman and Hogg 2010; Szabo and Huberman 2010) and Youtube (Pinto, Almeida, and Gonçalves 2013), early activity data consistently predicts future popularity with reasonable accuracy. In light of these results, we focus on the peeking variant of the problem in this paper.

Classification versus regression

In addition to how much data we look at, we must also specify what to predict. A number of studies have used regression formulations, predicting an item's exact final popularity: the number of retweets for a URL (Bakshy et al. 2011), votes on a Digg post (Lerman and Hogg 2010) or page views of a Youtube video (Szabo and Huberman 2010). However, we may often be more interested in popularity relative to other items rather than an exact estimate. For example, both marketers and platform owners may want to select 'up and coming' items to feature in the interface versus others[1].

These motivations lead nicely to a classification problem where the goal is to predict whether an item will be more popular then a certain percentage of other items. For instance, Romero et al. predict whether the number of adopters of a hashtag on Twitter will double, given a set of hashtags with the same number of initial adopters (Romero, Tan, and Ugander 2013). Cheng et al. generalize this formulation to show that predicting whether an item will double its popularity is equivalent to classifying whether an item becomes more popular than the median and study this question in the case of Facebook photos that received at least five adopters (Cheng et al. 2014). Besides the practical appeal of classifying popular items, classification is also a simpler task than predicting the actual number of adoptions (Bandari, Asur, and Huberman 2012), thus providing a favorable scenario for evaluating the limits of predictability of popularity. Therefore, we focus on the classification problem in this paper.

[1] Such featuring makes some items more salient than others and surely affects the final popularity of both featured and non-featured items; typically, formulations of the problem look at very small slices of early activity, which presumably minimizes these effects.

Study	Problem Formulation	Content	Structural	Early Adopters	Temporal
Bakshy et al. (2011)	Regression (no peeking)	n	–	**Y**	–
Martin et al. (2016)	Regression (no peeking)	n	–	**Y**	–
Szabo et al. (2010)	Regression	–	n	–	**Y**
Tsur et al. (2012)	Regression	**Y**	**Y**	–	**Y**
Pinto et al. (2013)	Regression	–	–	–	**Y**
Yu et al. (2015)	Regression	–	n	–	**Y**
Romero et al. (2013)	Classification ($k = \{1000, 2000\}, n = 50\%$)	–	**Y**	–	–
Cheng et al. (2014)	Classification ($k = 5, n = 50\%$)	n	**Y**	**Y**	**Y**
Lerman et al. (2008)	Classification ($k = 10, n = 80\%$)	–	**Y**	**Y**	–
Weng et al. (2013)	Classification ($k = 50, n = \{70, 80, 90\%\}$)	–	**Y**	n	–

Table 1: A taxonomy of problem formulations for popularity prediction, along with importance of feature categories. **Y** means that the features in the category were useful for prediction, n means they were tried but not as useful, and – that they were not studied. Most studies report temporal and structural features as important predictors.

Our problem: Peeking-based classification

Based on the above discussion, the general peeking-based classification problem can be stated as:

> **P1:** *Given a set of items and data about their early adoptions, which among them are more likely to become popular?*

This question has a broad range of formulations based on how we define the early activity period, how much activity we are allowed to poke at, and how we define *popular*. The early activity period may be defined in terms of time elapsed t since an item's introduction (Szabo and Huberman 2010), or in terms of a fixed number k of early adoptions (Romero, Tan, and Ugander 2013). Fixing the early activity period in terms of number of adoptions has the useful side-effect of filtering out items with less than k adoptions overall, both making the problem harder and eliminating unpopular (thus often uninteresting) items. For this reason, most past work on peeking-based classification defines early activity in terms of the number of adoptions k.

The popularity threshold for what is "popular" may also be set at different percentiles ($n\%$). Table 1 summarizes past work based on their choices of problem formulation and choice of (k, n). One common approach is to collect all items that have k or more adoptions, then peek into the first k adoptions and predict whether eventual popularity of items lies above or below the median (Cheng et al. 2014). We call this Balanced Classification since there are guaranteed to be an equal number of high and low popularity items. Another variation is to only consider the top-*n* percentile of items as high popularity (Lerman and Galstyan 2008), a formulation that is arguably better-aligned with most use cases around content promotion than Balanced Classification. However, it is also harder than Balanced Classification; for this reason, and to continue to align with prior work, we focus on Balanced Classification.

While restricting to items with k adoptions helps to level the playing field because it provides a set of comparably popular items to study, it ignores the *time taken* to reach k adoptions. Based on prior work, our suspicion is that in this formulation temporal features dominate the others. To control for this temporal signal, we later introduce a problem formulation where both k and t are fixed. That is, we collect all items that received exactly k adoptions in a given time period t, and then predict which of them would be in the top half of popular items. We call this the Temporally Matched Balanced Classification problem, and as we will see, changing the definition has a profound impact on the quality of the models.

Choosing features

We now turn to the selection of features for prediction. Part of the allure of modeling is that the features that prove important might give information about *why* some items become popular in ways that could be both practically and scientifically interesting. Features used in prior work can be broadly grouped into four main categories: content, structural, early adopters and temporal (Cheng et al. 2014). Table 1 shows which feature categories were used in prior studies, with cells in bold representing features that were reported to be useful for prediction. While all feature categories have been reported to be important contributors to prediction accuracy in at least some studies, temporal and structural features are frequently reported as important.

Temporal patterns of early adoption—how quickly the early adopters act—are a major predictor of popularity. Szabo and Huberman show that temporal features alone can predict future popularity reliably (Szabo and Huberman 2010). When information about the social network or its users is hard to obtain, utilizing temporal features can be fruitful, achieving error rates as low as 15% in a regression formulation (Pinto, Almeida, and Gonçalves 2013; Zhao et al. 2015). A natural next question is to ask how much these errors can be decreased by adding other features when we do have such information.

Features about the seed user and early resharers—collectively called early adopters—also matter. On Twitter, for example, the number of followers of the seed user and the fraction of her past tweets that received retweets increase the accuracy of predictions (Tsur and Rappoport 2012). Information about other early adopters is also useful for predicting photo cascades in Facebook (Cheng et al. 2014).

The structure of the underlying social network also has predictive power (Lerman and Galstyan 2008; Romero, Tan, and Ugander 2013; Cheng et al. 2014). However, these studies do not agree on the direction of effect of these features.

For instance, on Digg, low network density is connected with high popularity (Lerman and Galstyan 2008), but on Twitter, both very low and very high densities are positively correlated with popularity (Romero, Tan, and Ugander 2013). Their intuition is that a lower network density indicates that the item is capable of appealing to a general audience, while a higher network density indicates a tight-knit community supporting the item, both of which can be powerful drivers for an item's popularity.

Finally, while Tsur et al. report content features to be useful (Tsur and Rappoport 2012), most studies find content features to have little predictive power (Table 1). Even for domains such as songs or movies where item information is readily available, content features are not significantly associated with item popularity (Pachet and Sony 2012). Further, content features do not generalize well; it is hard to compute generalizable content features across different item domains. For these reasons, we do not consider content features in this work.

Features

Based on the above discussion, we use the following categories of features, with the aim of reproducing and extending the features used in past work (Cheng et al. 2014): temporal, structural, and early adopters. To these we add a set of novel features based on preference similarity between early adopters.

Temporal. These features have to do with the speed of adoptions during the early adoption period between the first and kth adoption. This leads to a set of features that focus on the rate of adoption:

- $time_i$: time between the initial adoption and the i^{th} adoption ($2 \leq i \leq k$). (Zhao et al. 2015; Maity et al. 2015; Weng, Menczer, and Ahn 2013)
- $time_{1...k/2}$: Mean time between adoptions for the first half (rounded down) of the adoptions.
- $time_{k/2...k}$: Mean time between adoptions for the last half (rounded up) of the adoptions.

Structural. These features have to do with the structure of the network around early adopters and can be broken down into two sub-categories: ego network features that relate the early adopters to their local networks, and subgraph features that consider only connections between the early adopters.
Early adopters' ego network features

- in_i: Indegree of the i^{th} early adopter ($2 \leq i \leq k$). This is a proxy for the number of people who may be exposed to an early adopter's activity. For undirected networks, this will simply be the degree, or the number of friends of an early adopter. (Bakshy et al. 2011; Zhao et al. 2015)
- $reach$: Number of nodes reachable in one step from the early adopters.
- $connections$: Number of edges from early adopters to the entire graph. (Romero, Tan, and Ugander 2013)

Early adopters' subgraph features

- $indegree_{sub}$: Mean indegree (friends or followers) for each node in the subgraph of early adopters. (Lerman and Galstyan 2008)
- $density_{sub}$: Number of edges in the subgraph of early adopters. (Romero, Tan, and Ugander 2013)
- cc_{sub}: Number of connected components in the subgraph of early adopters. (Romero, Tan, and Ugander 2013)
- $dist_{sub}$: Mean distance between connected nodes in the subgraph of early adopters. This is meant to measure how far the item has spread in the initial early adopters, similar to the cascade depth feature by Cheng et al.
- sub_in_i: Indegree of the i^{th} adopter on the subgraph ($1 \leq i \leq k$). (Lerman and Galstyan 2008)

Features of early adopters. These features capture information about early adopters, such as their popularity, seniority, or activity level, which might be proxies for their influence. They can be divided into two sub-categories: features of the first user to adopt an item (root), and features averaged over other early adopters (resharers).
Root features

- $activity_{root}$: Number of adoptions in the four weeks before the end of the early adoption period. This is similar to a measure used by Cheng et al. which measured the number of days a user was active. (Cheng et al. 2014; Petrovic, Osborne, and Lavrenko 2011; Yang and Counts 2010)
- age_{root}: Length of time the user has been registered on the social network.
- $popularity_{root}$: Number of friends or followers on the social network. (Lerman and Galstyan 2008; Tsur and Rappoport 2012)

Resharer features

- $activity_{resharer}$: Mean number of adoptions in the four weeks before the end of the early adoption period.
- $age_{resharer}$: Mean length of time the users have been registered on the social network.
- $popularity_{resharer}$: Mean number of friends or followers on the social network. (Tsur and Rappoport 2012)

Similarity To these previously tested features, we add features related to preference similarity between the early adopters. As with network density, our intuition is that similarity between early adopters may matter in two ways: high similarity may signify a niche item, or one that people with similar interests are likely to adopt, while low similarity might indicate an item that could appeal to a wide variety of people.

Similarity was computed using the Jaccard index of two users' adoptions that occurred before the end of the early adoption period of the item in question. We computed the median, mean and maximum of similarity between adopters because these give us an idea of the distribution of the affinity of the early adopters; we do not include users who had less than five adoptions before the item in question because they are likely to have little overlap. The features we extracted are:

Dataset	Last.fm	Twitter	Flickr	Goodreads
Number of users	437k	737k	183k	252k
Number of items	5.8M	64k	10.9M	1.3M
Number of adoptions	44M	2.7M	33M	28M
Mean adoptions	7.6	41.8	3.0	21.4
Median adoptions	1	1	1	1
Maximum adoptions	11062	82507	2762	88027

Table 2: Descriptive statstics for users, items, and adoptions in each dataset. We use *adoption* to mean loving a song on Last.fm, tweeting a URL on Twitter, favoriting a photo on Flickr, and rating a book on Goodreads. The average number of adoptions per item varies quite a bit, but the median popularity of 1 is consistent across datasets.

- sim_{count}: Number of similarities that could be computed between early adopters.
- sim_{mean}: Mean similarity between early adopters.
- sim_{med}: Median similarity between early adopters.
- sim_{max}: Maximum similarity between early adopters.

Data and Method

Datasets from four online social networks

We build models using data from four different online social platforms: Last.fm, Flickr, Goodreads and Twitter. These platforms span a broad range of online activity, including songs, photos, books and URLs; they also have a variety of user interfaces, use cases, and user populations. These variations reduce the risk of overfitting to properties of a particular social network.

- **Last.fm:** A music-focused social network where users can friend one another and love songs. We consider a dataset of 437k users and the songs they loved from their start date until February 2014 (Sharma and Cosley 2016).
- **Flickr:** A photo sharing website where users can friend one another and favorite photos. We use data collected over 104 days in 2006 and 2007 (Cha, Mislove, and Gummadi 2009).
- **Goodreads:** A book rating website where users can friend one another and rate books. The dataset consists of 252k users and their ratings before August 2010. Unlike the other sites, Goodreads users rate books; we consider any rating at or above 4 (out of 5) as an endorsement (adoption) of the book (Huang et al. 2012).
- **Twitter:** A social networking site where users can form directed edges with one another and broadcast *tweets*, messages no longer than 140 characters (as of 2010). The Twitter dataset consists of URLs tweeted by 737k users for three weeks of 2010 (Hodas and Lerman 2014).

All of these websites have an active social network, providing an activity feed that allows users to explore, like, and reshare the items that their friends shared. The Last.fm feed shows songs that friends have to listened to or *loved*, Flickr shows photos that friends have *favorited*, Goodreads shows books that friends have *rated*, and Twitter shows tweets with URLs that followees have *favorited* or *retweeted*. Thus, like past studies on online social networks such as Facebook, Twitter and Digg, we expect active peer influence processes that should make structural and early adopter features relevant.

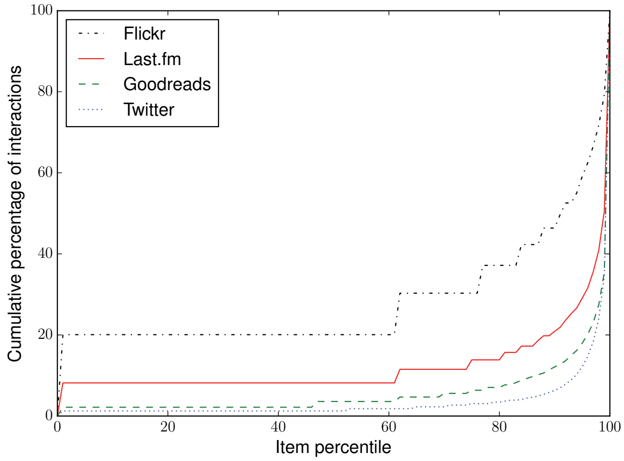

Figure 1: Cumulative percentage of adoptions by items for each dataset. Items on the x-axis are sorted by their popularity; the lines show a step pattern because multiple items may have the same number of adoptions. We observe a substantial skew in popularity. For example, the most popular 20% of items account for 60% of adoptions in Flickr and more than 90% of adoptions in other datasets.

Table 2 shows descriptive statistics about the datasets, all of which have more than 150k users and millions of items (with the exception of Twitter with 64k URLs). Twitter has the highest mean adoptions per item (41), followed by Goodreads (21). The maximum number of adoptions for an item also varies, from more than 80k in Twitter and Goodreads to 2.7k in Flickr. The median number of adoptions is consistent, however: at least half of the items have only 1 adoption. The skew in popularity distribution is better shown in Figure 1. The 20% of the most popular items account for over 60% of adoptions in Flickr and over 90% of the adoptions in the other three websites. On Twitter, the skew is extreme: over 81% adoptions are on 4% of items.

Classification methodology

We first operationalize the Balanced Classification formulation on these datasets. As a reminder, k is the number of early adoptions that we peek at for each item, and we pre-

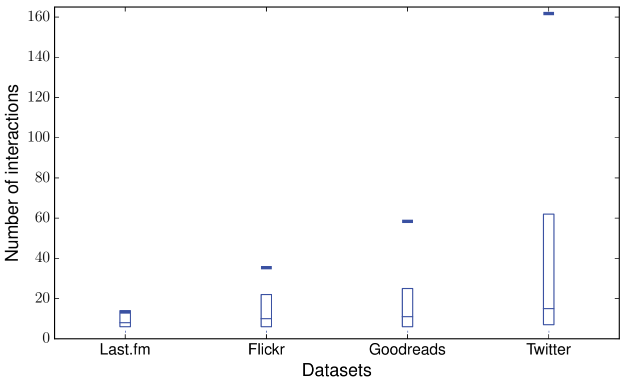

Figure 2: Boxplot showing the number of adoptions after 28 days (10 for Twitter) for items which have at least 5 adoptions. The bold partial line is the mean number of adoptions. Across datasets, most items receive less than 20 adoptions.

dict which of these items will end up more popular than the median item.

We measure the final popularity at a time T days after the first adoption of the item. To be consistent with prior work, we follow Cheng et al. and set $k = 5$ and $T = 28$ days for Last.fm, Flickr and Goodreads. Because the Twitter dataset is only three weeks long, we use a smaller $T = 10$ days. To avoid right-censoring, we include only items that had their first adoption at least T days before the last recorded timestamp in each dataset. The parameter k also acts as a filter, allowing only items with at least k adoptions. Figure 2 shows properties of the data thus constructed.

We classify items based on their popularity after T days, labeling those above the median 1 and others as 0. For each item, we extract features from the early adoption period, the time between the first and kth adoption. We use 5-fold cross validation to select the items that we train on, then use the trained model to predict final popularity of items in the test set. Since we use median popularity as the classification threshold, the test data has a roughly equal number of items in each class, allowing us to use accuracy as a reasonable evaluation metric. We tried several classification models using Weka (Hall et al. 2009), including logistic regression, random forests and support vector machines. Logistic regression models generally performed best, so we report results for those models unless otherwise specified.

Balanced classification

We start by comparing the predictive power of models using different sets of features across the four datasets on the Balanced Classification problem.

Temporal features dominate

Figure 3 shows the prediction accuracy of the models. Similar to prior work on Facebook that used peeking (Cheng et al. 2014), when using all features we are able to predict whether an item will be above the median popularity around three-fourths of the time: 73% for Goodreads, 75% for Flickr, 81% for Last.fm and 83% for Twitter.

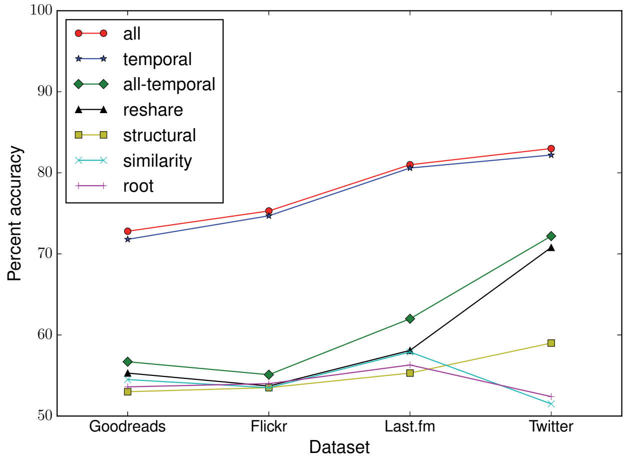

Figure 3: Accuracy for prediction models incorporating different categories of features. The y-axis starts at 50%, the baseline for a random classifier on the balanced formulation. On all datasets, temporal features are the most predictive, almost as accurate as using all available features.

Training models with individual feature categories shows that temporal features are by far most important. Across all four datasets, a model using only temporal features performs almost as well as the full model. The next best feature category, resharer features, is able to predict 71% on Twitter and less than 60% on the other three datasets. Even a model that uses *all* non-temporal features, denoted by the "all-temporal" line in Figure 3, is not very good. For Goodreads and Flickr, this model is not much better than a random classifier. For Last.fm and Twitter, accuracy for non-temporal features improves somewhat, but is still at least 10% worse than when including temporal features.

Even a single temporal feature can be more predictive than models constructed from all non-temporal features. Consider the feature $time_x$, which is the number of days for an item to receive x number of adoptions. At $x = 5 = k$, the feature $time_5$—time taken for an item to receive 5 adoptions—is the most predictive temporal feature for all datasets. A model based on this single feature achieves more than 70% accuracy on all datasets and accounts for nearly 97% of the accuracy of the full model for each dataset. While past work has highlighted the importance of temporal features as a whole (Szabo and Huberman 2010; Cheng et al. 2014), it is interesting to find that we may not even need multiple temporal features: a single measure is able to predict final popularity class label for items in all datasets.

Cross-domain prediction

The analysis in the previous section confirms past findings about the importance of temporal features across a range of websites. We now extend these results to show that temporal features are not only powerful, they are also general: models learnt on one item domain using temporal features are read-

Test \ Train	Last.fm	Flickr	Goodreads	Twitter
Using only temporal features				
Last.fm	**80.6**	80.7	78.0	80.0
Flickr	73.9	**74.7**	70.0	73.9
Goodreads	70.3	69.7	**71.9**	70.3
Twitter	82.7	82.3	79.7	**82.2**
Using only non-temporal features				
Last.fm	**62.1**	56.3	60.2	52.3
Flickr	53.0	**55.1**	51.8	48.2
Goodreads	56.0	52.1	**57.1**	50.6
Twitter	45.8	44.1	56.4	**73.4**

Table 3: Prediction accuracy for models trained on one dataset (columns) and tested on each dataset (rows). The diagonals report accuracy on the same dataset, while other cells report accuracy when the model is trained on one dataset and tested on another. The power of temporal features generalizes across domains: testing a model on any dataset, trained on any other dataset, loses no more than 5% accuracy compared to testing a model on the same dataset. For non-temporal features, prediction accuracy decreases substantially when applying models to other datasets.

ily transferable to others. In contrast, non-temporal features do not generalize well: even the direction of their effect is not consistent across domains. To show this, we train prediction models separately for each dataset, as before, then apply each model to every dataset.

Temporal features generalize

Table 3 shows the accuracy of models trained only on temporal features from one dataset and tested on all four. Reading across the rows shows that regardless of which social network a model was trained on, its accuracy on test data from another network remains within 5% of the accuracy on test data from the same network.

Such consistent prediction accuracy is impressive, especially because the median time to reach 5 adoptions varies, ranging from 1 day in Flickr to 15 days for Goodreads. This suggests that there are general temporal patterns that are associated with future popularity, at least across these particular networks.

Other features have inconsistent effects

The story is less rosy for non-temporal features. Table 3 shows the cross-domain prediction accuracy for models trained on all non-temporal features (in light of their low accuracy when taken individually, we combine all non-temporal features). Accuracies on the same dataset correspond to the "all-temporal" line in Figure 3; they are generally low and drop further when tested on a different dataset. In particular, models trained on other websites do poorly when tested on Twitter, with the Last.fm and Flickr models performing worse than a random guesser on Twitter data. Meanwhile, a model trained on Twitter is almost 10 percentage points worse than the Last.fm-trained model for predicting popularity on Last.fm.

Not only does prediction accuracy drop across websites, but fitting single-feature logistic regression models for each feature shows that for 12 of the 25 features, the coefficient term flips between being positive and negative across models fit on different datasets. Similar to the contrasting results found in prior work (Lerman and Hogg 2010; Romero, Tan, and Ugander 2013), we find that all measures of subgraph structural features of the early adopters, namely $indegree_{sub}$, $density_{sub}$, cc_{sub}, $dist_{sub}$ and sub_in_i (except for sub_in_1 and sub_in_4), can predict either higher or lower popularity depending on the dataset. For example, a higher $density_{sub}$—number of edges in the subgraph of early adopters—is associated with higher popularity on Flickr (β coefficient=0.04), whereas on Last.fm, a higher density is associated with lower popularity (β coefficient=-0.09). Features from the root, resharer and similarity categories show a similar dichotomous association with final item popularity.

Gaps between prediction and understanding

These results show that not only are non-temporal features weak predictors, the direction of their effect on popularity is inconsistent across different domains. Combining this with our observation that a single temporal heuristic is almost as good a predictor as the full model raises questions about what it is that popularity prediction models are predicting and how they contribute to our understanding of popularity.

Temporal features drive predictability

While our work may seem contrary to recent work that claims that early adopters and properties of their social network matter for prediction, many of their findings are consistent with our own. Most prior work that uses peeking finds that temporal features are a key predictor (Tsur and Rappoport 2012; Szabo and Huberman 2010; Pinto, Almeida, and Gonçalves 2013; Yu et al. 2015). Further, even though Cheng et al. conclude temporal and structural features are major predictors of cascade size, they report for predicting photos' popularity on Facebook, accuracy for temporal features alone (78%)is nearly as good as the full model (79.5%) (Cheng et al. 2014).

By holding modeling, feature selection and problem formulation consistent, we contribute to this literature by demonstrating the magnitude and generality of the predictive power of temporal features across a range of social networks. Having multiple networks also lets us show that, unlike temporal features, using non-temporal features does not generalize well to new contexts. These features might be useful for understanding the particulars of a given website, but it seems likely that they are capturing idiosyncrasies of that site rather than telling us something general about how items become popular in social networks.

Is cumulative advantage the whole story?

If non-temporal features are weakly predictive and not generalizable, and all that matters is the rate of initial adoption, then how do predictive exercises with peeking advance scientific understanding of what drives popularity? In other words, what does it mean when one claims that popularity is predictable once we know about initial adopters?

One answer is that early, rapid adoption is a signal of intrinsic features of an item that help to determine its popularity. Items with better content command a higher initial popularity, and thus the predictive power of early temporal features is simply a reflection of content quality or interestingness to the social network in question. Given increasing evidence from multiple domains that content features are at best weakly connected to an item's popularity (Salganik, Dodds, and Watts 2006; Pachet and Sony 2012; Martin et al. 2016), this seems unlikely to be the whole explanation.

Another explanation is that items that get attention early are more likely to be featured in the interface, via feeds, recommendations or ads; they might also be spread through external channels could drive up the rate of early adoption. Those would be interesting questions to explore. Still, whatever be the driving reasons, these models are telling us that once items achieve initial popularity, they are much more likely to become more popular in the future. This is simply a restatement of cumulative advantage, or the rich-get-richer phenomenon (Borghol et al. 2012).

Overall, though, we find that neither our results nor other work say much about why or how items become popular, except that items that share temporal patterns of popular items early on tend to be the ones that are more popular in the future, and that making popularity salient and ordering items by popularity can increase this effect (Salganik, Dodds, and Watts 2006). While such predictions are practically useful for promoting content, they are not so useful for informing creation of new content or assessing its value, nor for understanding the mechanisms by which items become popular.

Temporally matched balanced classification

In this section, we give a problem formulation that lessens the importance of temporal features by conditioning on the average rate of adoption. That is, instead of considering all items with k adoptions, we consider items with k adoptions within about the same amount of time. Given the dominance of cumulative advantage, such a formulation would be better suited for future research in understanding how items become popular, as gains in accuracy will likely shed light on attributes of early adopters, items, and networks that affect their final popularity.

k-t problem formulation

We call this formulation Temporally Matched Balanced Classification, or a *k-t* formulation of the problem:

> **P2:** *Among items with exactly k adoptions at the end of a fixed time period t, which ones would be higher than the median popularity at a later time T?*

To do this, for each dataset we filtered items to those that had exactly k adoptions in t days. We extracted features of these items as previously described, adding a new temporal feature for each day in t:

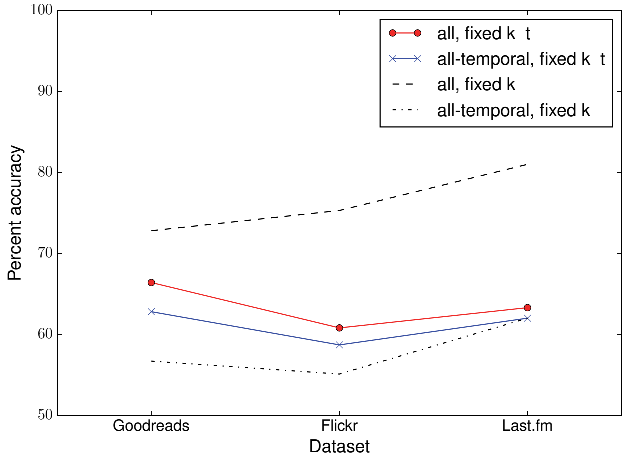

Figure 4: Percent accuracy for fixed t & k using all features and non-temporal features, and for fixed k with all features and non-temporal features. $k = 5$, $T = 28$ days for all; $t = 15$ days for Goodreads, $t = 1$ days for Flickr, and $t = 7$ days for Last.fm. Fixing t reduces accuracy substantially compared to when t is not fixed. As expected when controlling for time, non-temporal features now provide most of the explanatory power.

- $adoptions_i$: Number of adoptions on day i of the early adopter period. (Szabo and Huberman 2010; Tsur and Rappoport 2012; Pinto, Almeida, and Gonçalves 2013)

As before, we choose $k = 5$ and $T = 28$ days. For each dataset, we set t to be the median time it took an item to reach five adoptions: $t = 15$ for Goodreads, $t = 7$ for Last.fm, and $t = 1$ for Flickr. We exclude Twitter due to a lack of data when we filter for both k and t. We again do 5-fold cross-validation, predicting if each item would be above or below the final median popularity after T days.

Figure 4 shows the results. As we hoped, non-temporal features now provide most of the explanatory power in the full model. Further, comparing the all-temporal series with fixed k and t to the one with only fixed k shows that the absolute accuracy of non-temporal features increases in this formulation. This suggests that de-emphasizing temporal features in prediction might in fact improve our understanding of other features that drive popularity.

Our understanding, however, is limited: even conditioning on a single temporal feature makes for a much harder problem, with the overall prediction accuracy below 65% for all datasets even when using all features. There is clearly much room for improvement.

Discussion and Conclusion

Using multiple problem formulations, we show that temporal features matter the most in predicting the popularity of items given data about initial adopters and our current ability to build explanatory features of those adopters and their networks. Using datasets from a variety of social networks, we show that temporal features are not only better at predicting popularity than all other features combined, but that they readily generalize to new contexts. When we discount temporal phenomena by removing temporal features or adjusting the problem formulation, accuracy decreases substantially.

From a practical point of view, these results provide empirical support for a promising approach where only temporal features are used to predict future popularity (Szabo and Huberman 2010; Zhao et al. 2015) because the drop in accuracy by casting aside non-temporal features is generally small. Maybe creative feature engineering is not worth the effort for the Balanced Classification task. This way of looking at the problem resonates a bit with the Netflix prize, where most of the learners that were folded into the winning model were never implemented in Netflix's actual algorithm, in part because the cost of computing and managing those learners was not worth the incremental gains (Amatriain and Basil 2012).

Although less valuable than temporal features, the non-temporal features examined so far do have some predictive power on their own. This might be useful when temporal information is unavailable: for example, for very new items (Borghol et al. 2012), or for external observers or datasets where timestamps are unavailable (Cosley et al. 2010). Encouragingly, non-temporal features increase in accuracy a little on the *k-t* formulation compared to the fixed-k balanced classification problem, suggesting that making time less salient might allow other factors to become more visible and modelable.

Using *k-t* models could also bend time to our advantage. Comparing the overall performance and predictive features in models with smaller versus larger t might highlight item, adopter, and network characteristics that predict faster adoption (and eventual popularity). Another way to frame this intuition is that instead of predicting eventual popularity, we should try to predict initial adoption speed.

Deeper thinking about the context of sharing might also be useful. Algorithmic and interface factors, for instance, have been shown to create cumulative advantage effects; it would be interesting to look more deeply into how system features might influence adoption behaviors. Likewise, diffusion models tend to focus attention on sharers rather than receivers of information—but those receivers' preferences, goals and attention budgets likely shape their adoption behaviors (Sharma and Cosley 2015). Thus, consideration of audience-based features might be a way forward.

Most generally, we encourage research in this area to go beyond the low-hanging fruit of time. For building better theories of diffusion, maximizing accuracy with temporal information may act both as a crutch that makes the problem too easy, and as a blindfold that makes it hard to examine what drives those rapid adoptions that predict eventual popularity.

Acknowledgments

This work was supported by the National Science Foundation under grants IIS 0910664 and IIS 1422484, and by a grant from Google for computational resources.

References

Amatriain, X., and Basil, J. 2012. Netflix recommendations: Beyond the 5 stars. *http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html*.

Asur, S.; Huberman, B.; et al. 2010. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, 492–499.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*.

Bandari, R.; Asur, S.; and Huberman, B. A. 2012. The pulse of news in social media: Forecasting popularity. In *Sixth International AAAI Conference on Weblogs and Social Media*, 26–33.

Berger, J., and Milkman, K. L. 2012. What makes online content viral? *Journal of marketing research* 49(2):192–205.

Berger, J. 2013. *Contagious: Why things catch on*. Simon and Schuster.

Borghol, Y.; Ardon, S.; Carlsson, N.; Eager, D.; and Mahanti, A. 2012. The untold story of the clones: content-agnostic factors that impact youtube video popularity. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1186–1194.

Cha, M.; Mislove, A.; and Gummadi, K. P. 2009. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, 721–730.

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, 925–936.

Cosley, D.; Huttenlocher, D. P.; Kleinberg, J. M.; Lan, X.; and Suri, S. 2010. Sequential influence models in social networks. *Fourth International AAAI Conference on Weblogs and Social Media* 10:26.

Frank, R. H., and Cook, P. J. 2010. *The winner-take-all society: Why the few at the top get so much more than the rest of us*. Random House.

Gladwell, M. 2006a. The formula. *The New Yorker*.

Gladwell, M. 2006b. *The tipping point: How little things can make a big difference*. Little, Brown.

Goel, S.; Hofman, J. M.; Lahaie, S.; Pennock, D. M.; and Watts, D. J. 2010. Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences* 107(41):17486–17490.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter* 11(1):10–18.

Hodas, N. O., and Lerman, K. 2014. The simple rules of social contagion. *Scientific reports* 4.

Huang, J.; Cheng, X.-Q.; Shen, H.-W.; Zhou, T.; and Jin, X. 2012. Exploring social influence via posterior effect of word-of-mouth recommendations. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 573–582.

Jenders, M.; Kasneci, G.; and Naumann, F. 2013. Analyzing and predicting viral tweets. In *22nd international conference on World Wide Web*, 657–664.

Kupavskii, A.; Umnov, A.; Gusev, G.; and Serdyukov, P. 2013. Predicting the audience size of a tweet. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Lerman, K., and Galstyan, A. 2008. Analysis of social voting patterns on digg. In *Proceedings of the first workshop on Online social networks*, 7–12.

Lerman, K., and Hogg, T. 2010. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on World wide web*, 621–630.

Maity, S. K.; Gupta, A.; Goyal, P.; and Mukherjee, A. 2015. A stratified learning approach for predicting the popularity of twitter idioms. In *Ninth International AAAI Conference on Web and Social Media*.

Martin, T.; Hofman, J. M.; Sharma, A.; Anderson, A.; and Watts, D. J. 2016. Limits to prediction: Predicting success in complex social systems. In *Proceedings of the 25th international conference on World wide web*.

Pachet, F., and Sony, C. 2012. Hit song science. *Music Data Mining* 305–26.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2011. Rt to win! predicting message propagation in twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Pinto, H.; Almeida, J. M.; and Gonçalves, M. A. 2013. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*.

Romero, D. M.; Tan, C.; and Ugander, J. 2013. On the interplay between social and topical structure. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*.

Sharma, A., and Cosley, D. 2015. Studying and modeling the connection between people's preferences and content sharing. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1246–1257.

Sharma, A., and Cosley, D. 2016. Distinguishing between personal preferences and social influence in online activity feeds. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1091–1103.

Simonoff, J. S., and Sparrow, I. R. 2000. Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance* 13(3):15–24.

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

Tsur, O., and Rappoport, A. 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 643–652.

Watts, D. J. 2011. *Everything is obvious:* Once you know the answer*. Crown Business.

Weng, L.; Menczer, F.; and Ahn, Y.-Y. 2013. Virality prediction and community structure in social networks. *Scientific reports* 3.

Yang, J., and Counts, S. 2010. Predicting the speed, scale, and range of information diffusion in twitter. *Fourth International AAAI Conference on Weblogs and Social Media* 10:355–358.

Yu, L.; Cui, P.; Wang, F.; Song, C.; and Yang, S. 2015. From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics. *IEEE International Conference on Data Mining*.

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1513–1522.