Popularity Dynamics of Foursquare Micro-Reviews

Marisa Vasconcelos, Jussara Almeida, Marcos Gonçalves,
Daniel Souza, Guilherme Gomes
{marisav,jussara,mgoncalv,daniel.reis,gcm.gomes}@dcc.ufmg.br
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

ABSTRACT

Foursquare, the currently most popular location-based social network, allows users not only to share the places (venues) they visit but also post micro-reviews (tips) about their previous experiences at specific venues as well as "like" previously posted tips. The number of "likes" a tip receives ultimately reflects its popularity among users, providing valuable feedback to venue owners and other users.

In this paper, we provide an extensive analysis of the popularity dynamics of Foursquare tips using a large dataset containing over 10 million tips and 9 million likes posted by over 13,5 million users. Our results show that, unlike other types of online content such as news and photos, Foursquare tips experience very slow popularity evolution, attracting user likes through longer periods of time. Moreover, we find that the social network of the user who posted the tip plays an important role on the tip popularity throughout its lifetime, but particularly at earlier periods after posting time. We also find that most tips experience their daily popularity peaks within the first month in the system, although most of their likes are received after the peak. Moreover, compared to other types of online content (e.g., videos), we observe a weaker presence of the rich-get-richer effect in our data, demonstrating a lower correlation between the early and late popularities. Finally, we evaluate the stability of the tip popularity ranking over time, assessing to which extent the current popularity ranking of a set of tips can be used to predict their popularity ranking at a future time. To that end, we compare a prediction approach based solely on the current popularity ranking against a method that exploits a linear regression model using a multidimensional set of predictors as input. Our results show that use of the richer set of features can indeed improve the prediction accuracy, provided that enough data is available to train the regression model.

COSN'14, October 1–2, 2014, Dublin, Ireland.

http://dx.doi.org/10.1145/2660460.2660484.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services; J.4 [**Computer Applications**]: Social and behavioral sciences

Keywords

Popularity dynamics; online micro-reviews; location-based social networks

1. INTRODUCTION

Understanding the popularity dynamics of online content, particularly user generated content (UGC), is quite a challenge due to the various factors that might affect how the popularity of a particular piece of content (here referred to as an object) evolves over time. Moreover, the processes that govern UGC popularity evolution may vary greatly depending not only on the type of content but also on characteristics of the particular application where the content is shared. For example, mechanisms employed by the application, such as search and recommendation, social links among users, and even elements of the application that might favor the visibility of some objects over the others, may affect how content popularity evolves.

We here analyze the popularity dynamics of an increasingly popular type of UGC, namely Foursquare micro-reviews, also called *tips*, estimating the popularity of a tip at a certain time t by the number of likes it received from posting time until t.

A study of how the popularity of a tip evolves over time allows us to compare tips against other types of content whose popularity and dissemination dynamics have already been studied, such as videos [28, 31], photos [6, 34], and tweets and news articles [32]. Tips have inherent characteristics that distinguish them from these other types of content and that might impact their popularity evolution. For example, tips are associated with specific venues, and thus are visible to all users that visit the venue, including those that are drawn to it by other reasons (e.g., other tips). Also, tips usually contain opinions that might interest others for much longer periods of time than news and tweets. Thus, tips may remain live in the system, attracting attention (and likes), for longer periods.

The present effort also complements prior studies on the automatic assessment of the helpfulness (or quality) of online reviews, which focused mainly on traditional (longer) reviews, often exploiting textual features [17, 38]. Unlike such reviews, tips are more concise (constrained to 200 characters), often containing more subjective and informal content. Thus, attributes used by existing solutions, particularly those related to the textual content, may not be adequate for assessing the popularity of shorter reviews. Moreover, we are not aware of any prior study that analyzed the temporal popularity evolution of online reviews.

The study of tip popularity dynamics (as of any other type of content) can also provide valuable insights into improvements to the system. For example, it can guide the future design of tip popularity prediction methods [35], which in turn, can benefit various other services, including content filtering and recommendation, as well as more cost-effective marketing strategies. In the particular context of Foursquare tips, such predictions can benefit both users and venue owners as they can react quickly to opinions that may have a greater impact on decision making. For example, business owners are able to more quickly identify (and fix) aspects of their services or products that may affect revenues most.

In this context, we here provide an extensive analysis of the popularity dynamics of Foursquare tips. Using a large dataset containing over 10 million tips and 9 million likes posted by over 13,5 million users, we characterize how the popularity of different sets of tips evolves over time, and how it is affected by the social network of the user who posted the tip (its author). We observe that tips experience a very slow popularity evolution, compared to other types of UGC. While news articles acquire most of their comments within the first day of publication [32] and Flickr photos obtain half of their views within two days [34], tips take a couple of months to attract their likes. The social network of the tip's author has an important influence on the tip popularity throughout its lifetime, but especially in earlier periods after posting. For example, 62% of the likes received by the most popular tips during the first hour come from the social network of the user who posted them. This fraction is even larger for the less popular tips.

We also analyze tip popularity at and around the daily peak, and assess to which extent the rich-get-richer phenomenon impacts the popularity evolution of tips. We find that most tips experience their daily popularity peak within a month in the system. Yet, these peaks usually correspond to a small fraction of the total popularity, as most likes are received after the daily peak. Compared to YouTube videos [4], we observe a weaker presence of the rich-get-richer phenomenon in the popularity evolution of tips, suggesting that other factors, but the current popularity, may significantly impact the tip's future popularity.

Finally, we assess to which extent the future relative popularity of a set of Foursquare tips can be *predicted* based only on their popularity ranking at the prediction time, or, in other words, to which extent the tip popularity ranking remains stable over time. To that end, we compare two prediction strategies: one based solely on the current popularity ranking, and one that exploits a regression model and a much richer and multidimensional set of features, capturing aspects related to the user who posted the tip, the venue where it was posted, and its content. Our experimental results indicate that these features can improve the prediction accuracy, given that enough training data is available.

The rest of this paper is as follows. We review related work in Section 2 and describe our Foursquare dataset in Section 3. We analyze the dynamics of tip popularity in Section 4 and tackle the popularity ranking prediction problem in Section 5. Section 6 offers conclusions and directions for future work.

2. RELATED WORK

Our work is focused on analyzing the popularity evolution of Foursquare tips, estimated by the number of likes received. Previous related efforts can be grouped into: analyses of online content popularity, and methods to assess the helpfulness of online reviews.

Online Content Popularity. A number of studies on popularity dynamics were conducted analyzing the role of the social networks in the spread of news, videos [7, 20, 4], images [6] and tweets [20, 36]. Crane and Sornette [7] described four classes (memoryless, viral, quality and junk) of YouTube videos characterized by how their popularity evolves over time. The authors defined these classes according to the degree of influence of endogenous user interactions and external events. In contrast, Yang and Leskovec [36] proposed a clustering algorithm to classify the temporal evolution patterns of online content popularity, finding six "curves" that explain the popularity dynamics of tweets and news documents.

Lerman and Gosh [20] performed an empirical study to measure how popular news spread on Digg and Twitter. They observed that the number of votes and retweets accumulated by stories on both sites increases quickly within a short period of time and saturates after a day. In contrast, Cha et al. [6] showed that popular photos on Flickr, with popularity estimated by the number of favorite marks, spread neither widely nor rapidly through the network, contrary to the viral marketing intuition. Complementarily, Borghol et al. [4] assessed the impact of content-agnostic factors on the popularity of YouTube videos. They focused on groups of videos that have the same content (clones), finding a strong linear "rich-get-richer" behavior with the number of previous views as the most important factor.

Other studies have addressed the prediction of popularity of online content [1, 14, 28, 32]. Bandari et al. [1] and Hong et al. [14] exploited textual features extracted from messages (e.g., hashtags or URLs) or the topic of the message, and user related features to predict popularity of news and tweets. Tatar et.al [32] modeled the problem of predicting the popularity of a news article based on user comments as a ranking problem. Pinto et al. [28] proposed a multivariate regression model to predict the long-term popularity of YouTube videos based on measurements of user accesses during an early monitoring period. In [23], the authors proposed a unifying model for popularity evolution of blogs and tweets, showing that it can be used for tail-part forecasts.

Our current effort complements these prior studies by focusing on an inherently different type of content. Unlike news, videos and tweets, tips are associated with specific venues, and tend to be less ephemeral (particularly compared to news and tweets), as they remain associated with the venue (and thus visible to users) for a longer time. Thus, the analysis of tip popularity dynamics may lead to new insights. Also, towards analyzing the stability of popularity ranking over time, we tackle a different prediction task. While most prior efforts aim at predicting the future popularity of a given piece of content, we here explore strategies to predict the future popularity *ranking* of a set of tips.

Quality of Online Reviews. Most previous efforts to automatically assess the helpfulness or quality of online reviews

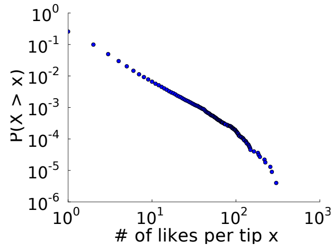


Figure 1: Distribution of Number of Likes per Tip

employ classification or regression-based models. For example, Kim et al. [17] used Support Vector regression (SVR) to rank reviews according to their helpfulness, exploiting features such as the length and the unigrams of a review and the reviewers' ratings. Mahony et al. [26] proposed a classification-based system to recommend the most helpful hotel reviews in Trip Advisor using features related to the user reviewing history and the scores previously assigned to the hotels. Zhang et al. [38], in turn, found that syntactic features (e.g., number of nouns, comparatives and modal verbs) extracted from the text reviews are the most effective predictors for SVR and linear regression to predict the utility of Amazon reviews. Ghose and Ipeirotis [13] applied a Random Forests classifier on a variety of textual features to predict if an Amazon product review is helpful or not. Hong et al. [15] built a binary helpfulness based system to classify Amazon reviews using textual features and features related to user preferences, and used this classification to rank product reviews. Finally, Momeni et al. [24] developed a "usefulness" classifier for predicting useful comments on YouTube and Flickr based on textual features as well as features that describe the author's posting and social behavior.

These prior studies focused on longer reviews, often exploiting textual features and, in some cases, aiming at a binary classification of reviews (helpful or not). Instead, we here tackle the ranking of tips based on the predicted number of likes. Tips have length constraints which lead users to write reviews using non-standard textual artifacts and informal language [3]. Thus, textual features often exploited are not adequate in our context. Moreover, previous work has not addressed how the helpfulness as perceived by users (or popularity) of the reviews evolve over time, as we do here.

The only prior study of tip popularity is a recent work of ours [35] which proposed regression and classification methods to predict, at posting time, the *popularity level* (high or low) of a given tip at a future time. We here greatly extend this work by: (1) providing an extensive analysis of tip popularity dynamics, and (2) tackling a different prediction task: the *ranking* of a set of tips based on their predicted popularity. Ranking and classification tasks support different applications. For example, tip ranking supports filtering and recommendation at a finer granularity (as opposed to 2 popularity levels) which is useful to users and venue owners.

3. FOURSQUARE DATASET

Our experiments are performed on a dataset consisting of more than 10 million tips posted by 13,5 million users at almost 16 million different venues. This dataset was crawled from Foursquare using the system's API from August to October 2011.

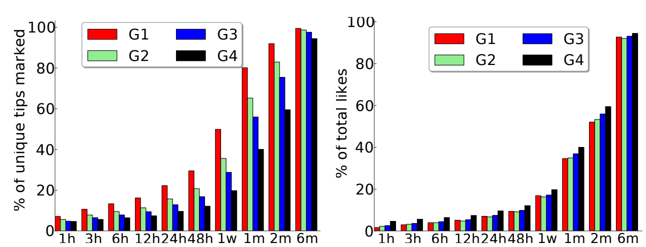
Figure 1 shows the complementary cumulative distribution of the number of likes received by each tip. The distribution is highly skewed, and only 34% of the tips received at least one like. As discussed in [35], this distribution, as the distributions of numbers of tips per user, likes per user, and tips per venue, are heavy tailed.

For the sake of analyzing tip popularity dynamics, we group tips with at least one like by breaking their popularity distribution into 10 slices, each one containing tips whose popularity fall into a certain range of the distribution[1]. For example, slice 0-10% contains the top-10% most popular tips, while slice 10%-20% contains the tips whose popularities fall between the 10^{th} and 20^{th} percentile of the popularity distribution. This partitioning is the same used in [34] for analyzing Flickr photos, since it is more balanced and less biased towards the more popular tips. Table 1 shows the number tips as well as total number of likes per slice.

Table 1: Distribution of Likes for Groups of Tips

Slice	# of Tips	Total # of Likes	% Social Likes	Group
0-10%	23,746	202,804	30.8%	G_1
10-20%	23,746	72,824	48.4%	G_2
20-30%	23,746	47,492	49.0%	G_3
30-40%	23,746	47,492	49.0%	G_3
40-50%	23,746	24,163	48.2%	G_4
50-60%	23,746	23,746	49.1%	G_4
60-70%	23,746	23,746	48.5%	G_4
70-80%	23,746	23,746	48.2%	G_4
80-90%	23,746	23,746	48.5%	G_4
90-100%	23,750	23,750	48.4%	G_4

We also examine the fraction of likes coming from the social network (friends and followers) of the user who posted the tip (i.e., the tip's author). Table 1 shows the percentages of likes coming from the social network, referred to as *social likes*, for tips in each slice. We note that for all slices but the first one, almost half of the likes received by tips come from the user's social network, highlighting the importance of friends and followers to the popularity of those tips. In contrast, for the most popular tips, the fraction of social likes is smaller (31%), suggesting that most likes probably come from venue visitors. We further analyze the importance of the social network to tip popularity in Section 4.2.



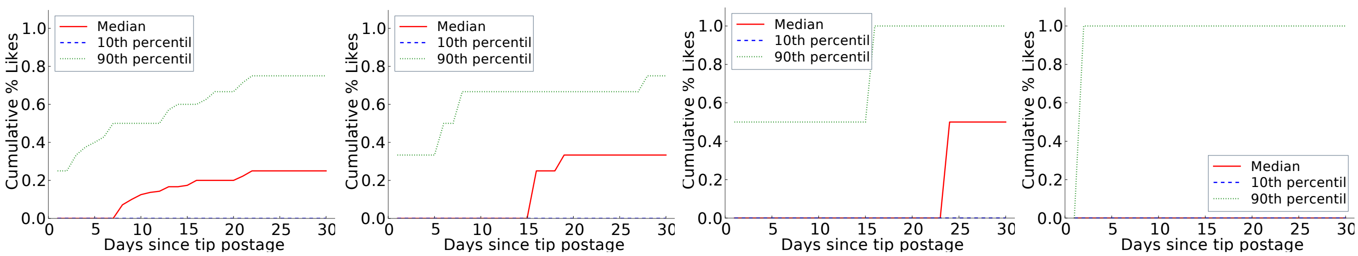
(a) Fraction of tips that received at least one like

(b) Fraction of total likes

Figure 2: Distribution of Tip Popularity over Time

We aggregate the slices into 4 major groups, as shown in Table 1. Groups 3 and 4 contain tips that received, on average, 2 and 1 likes, respectively. We analyze tip popularity

[1]Note that we exclude tips with no likes from these slices.



(a) G1 (b) G2 (c) G3 (d) G4

Figure 3: Distribution of Percentage of Likes Received During the First Month after Posting Time

separately for each slice. However, as the same conclusions hold for tips in different slices of the same group, we present results for each group only.

4. DYNAMICS OF TIP POPULARITY

In this section, we analyze the dynamics of tip popularity in Foursquare. We start by discussing how the number of likes of a tip evolves over time (Section 4.1), and how it is affected by the social network of the tip's author (Section 4.2). We then analyze tip popularity at and around the peak (Section 4.3), and assess to which extent the rich-get-richer phenomenon is present in the popularity evolution of tips (Section 4.4).

4.1 Popularity Evolution

We start by analyzing how the popularity of tips in each group of slices defined in Table 1 evolves over time. We focus on the first six months after the tip is posted. Figure 2a plots the fraction of unique tips in each group that received at least one like within the first x hours (h), week (w) or months (m) after posting time. We observe that within the first 48 hours, 29% of the tips in the most popular group (G1) received at least one like, while in one and two months this fraction grows up to 80% and 92%, respectively. That is, 20% of the top-10% most popular tips take more than one month to attract their first likes. This slow popularity evolution is even more clear for tips in the other (less popular) groups. Figure 2b shows the cumulative fraction of the total number of likes (as observed in our dataset) received by tips in each group over time. Note that, for all four groups, between 41% and 48% of the likes are received *after* 2 months since posting time.

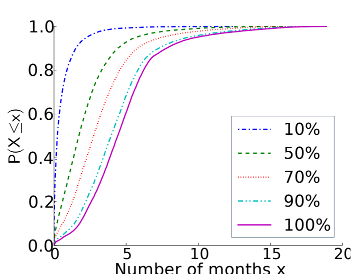


Figure 4: Distribution of time until x% of total likes are received for the most popular tips (G1)

Thus, in general, tips tend to live long in the system, presenting a gradual increase of interest. Indeed, tip popularity evolves much more slowly compared to other types of content, even for tips that end up becoming very popular. For example, news articles have a very short lifespan [32] acquiring all comments within the first day of publication, while a large fraction of views of Flickr photos are generated within the first two days after upload[34]. In contrast, we here find a significant fraction of tips that can take quite months to attract likes and become popular. This longer lifecycle was also observed in the acquisition of *fans* by Flickr photos [6].

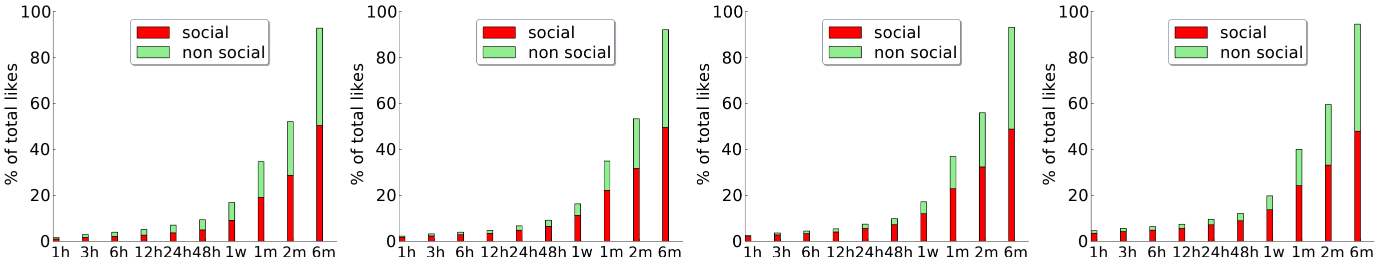
We further analyze the popularity evolution of tips in each group by showing in Figure 3 the curves of the 10^{th} and 90^{th} percentiles as well as the median of number of likes over time during the first one month since the tip was posted. For all groups, the 10^{th} percentile curve is equal to zero through the whole period, implying 10% of the tips in each group did not receive any like within the first month in the system. Around half of the most popular tips (G1) starts receiving likes after 7 days since posting time, achieving only 20% of the total likes after a month. For the second most popular group (G2), we note half of the tips start receiving likes after 15 days while tips in group G3 and G4 take more than 20 and 30 days, respectively, to start attracting likes.

We also analyze the amount of time it takes for a tip to receive at least X% of their total likes, for X equal to 10, 50, 70, 90 and 100%. Figure 4 shows those distributions for the most popular tips (G1). Note that 57% of the tips in this group take at least 2 (3) months to reach 50% (70%) of its total observed popularity. In sum, many tips do take a few months to attract likes, even those that end up being the most popular ones.

4.2 The Role of the Social Network

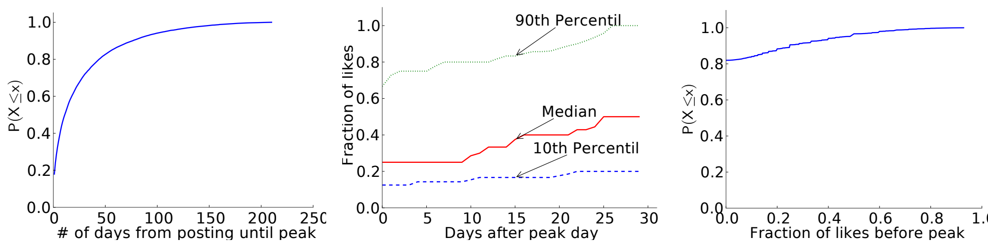
The popularity evolution of a tip is directly related to how users find the tip: either by visiting the venue page or through activity notifications from their friends and followees. Thus, the number of likes received by a tip depends on a combination of its visibility and interest by the social network of the tip's author and by others.

In this section, we discuss the role of the social network of the tip's author on its popularity evolution. To that end, we revisit Figure 2b by separating likes coming from the author's social network (social likes) and likes coming from other users (non-social likes). Figure 5 shows the cumulative fraction of likes, in both categories, for tips in each group. Note that the author's social network has an important influence on the popularity of a tip throughout its lifetime: at least half of all likes received in any period of time (up to 6 months since posting) come from the author's social net-



(a) G1 (b) G2 (c) G3 (d) G4

Figure 5: Social vs. Non Social Likes: Distribution of Percentage of Likes Received over Time



(a) Time Until Peak (b) % Likes At/After Peak (c) % Likes Before Peak

Figure 6: Cumulative Distributions of Popularity Peak for Most Popular Tips (G1)

work, for tips in *all* four groups. This fraction is higher in the earlier periods after posting time, and tends to decrease with time as the tip becomes visible to other users (e.g., venue visitors). For example, the social likes correspond to 62% of all likes received by the most popular tips (G1) in the 1^{st} hour since posting time, decreasing to 54% after 6 hours. Interestingly, the social network seems to have an even more important role for the least popular tips. For example, for tips in G2, G3 and G4, the social likes correspond to more than 70% of all likes received by a tip in the first week in the system.

These results indicate that the social network of a tip's author may be responsible for boosting its popularity, particularly during early periods after posting. As consequence, they also suggest that it might be possible for a recently posted tip to become more popular than other tips that had already attracted many likes and thus gained visibility in the system.

4.3 Popularity Peak

We further analyze tip popularity evolution by focusing on the popularity peak. Considering the daily popularity time series of each tip, we define the peak k_{p_i} of tip p_i as the largest number of likes received by p_i on a single day. We then compute the time (in number of days) it takes for p_i to reach is popularity peak[2]. We also measure the fraction of the total likes p_i received at, *before* and *after* the peak. For this analysis we focus on the most popular tips (G1).

Figure 6a shows the cumulative distribution of the time until the popularity peak. Around 18% of the tips experience its popularity peak one day after posting time, and around 72% of the tips reach their popularity peak within a month since posting. This implies that most tips do not take too long (less than a month) to reach its daily popularity peak. Yet, we observe that, for many tips, this peak represents only a small fraction of the total observed popularity. This is illustrated in Figure 6b, which presents the cumulative distributions of the median, 10^{th} and 90^{th} percentiles of the fraction of likes received at and after the peak day. As a complement, Figure 6c shows the cumulative distribution of the fraction of likes received *before* the peak day. Like observed for other types of online content (e.g., videos and news [7, 28, 32]), some tips do experience heavy bursts of popularity on the peak day: for 10% of the tips, the daily peak corresponds to at least 67% of their total popularity (see 90^{th} percentile curve in Figure 6b).

However, for half of the tips (median curve), the peak corresponds to only 25% of all likes. Moreover, Figure 6c shows that most tips (82%) receives their first like in the peak day, and only a very small fraction of the tips (3.3%) receive more than 50% of the likes before the peak day. Thus, a large fraction of tips receive most of their likes *after* the peak day, suggesting, once again, that tips experience a slow popularity evolution.

Contrasting our findings with the acquisition of fans by Flickr photos [6], we observe that both fans and likes are acquired after a longer period of time after posting/upload, compared to, for example, tweets. Also, as in [6], we do not observe an exponential growth on popularity as suggested by existing models of information diffusion [33]. However, comparing our results (particularly Figure 3), with similar ones presented in [6], we find that tip popularity seems to increase even more slowly than photo fan acquisition. For example, we do not observe a period of steady linear popularity growth during the first month, as observed for photos.

[2] In case of ties, we pick the first day with k_{p_i} likes.

4.4 The Rich-Get-Richer Phenomenon

Most online systems offer their users the option to see different pieces of content (or *objects*) sorted by their posting dates or by some estimate of their popularity. The adopted strategy may have a direct impact on the visibility of different objects. For example, by displaying objects sorted in decreasing order of popularity, a website may contribute to further increasing the popularity of an object that is already very popular, a phenomenon that is known as *rich-get-richer* [2]. Indeed, prior work has already suggested that popularity of some types of online content (e.g., video) evolves according to this phenomenon [4, 31].

Foursquare tips may be sorted by the number of likes (in increasing/decreasing order) or by posting time, but only the former is available in the mobile application. Thus, we here assess to which extent the rich-get-richer phenomenon can explain tip popularity evolution.

The rich-get-richer, or preferential attachment, models define that the probability of a tip p_i experiencing an increase in popularity is directly proportional to p_i's current popularity [2]. As in [4], we consider a model where the probability that a tip p_i with l_{p_i} likes receives a new like is a power law, i.e., $Prob(p_i) \propto l_{p_i}^{\alpha}$.

We analyze the rich-get-richer effect using a univariate linear regression to observe the impact of the number of likes of a tip after a monitoring time t_r (predictor variable) in the total number of likes of the tip at target time $t_r + \delta$ (response variable), using log-transformed data. The case of α=1 corresponds to a linear preferential selection [2], and $\alpha > 1$ implies in a case where the rich gets much richer with time. The sublinear case ($\alpha < 1$) results in a (stretched) exponential popularity distribution, which reflects a much weaker presence of the rich-get-richer effect [19]. We perform this analysis separately for tips in each group as well as for all tips.

Table 2 shows the coefficients α (along with 95% confidence intervals) and the coefficients of determination R^2 of the univariate regressions performed using various predictor and response variables, for tips in G1 as well as for all tips. For all considered cases, we find $\alpha < 1$, which indicates an exponential popularity evolution that could result in a much more even popularity distribution than suggested by the pure (linear) rich-get-richer dynamics. This has also been observed for a set of different YouTube videos [4], although the values of α found in that case (0.93 on average) are much larger than those we observed in all considered scenarios. This suggests that the rich-get-richer effect might be weaker in Foursquare tips than in YouTube videos, even considering all tips jointly. This also implies that other factors might strongly impact tip popularity. Indeed, as discussed in Section 4.2, the social network of the tip's author is responsible for a significant fraction of the likes received by the tip, and thus might contribute to reduce the impact of the rich-get-richer effect.

The univariate regression model has also been proposed as a means to predict the future popularity of YouTube videos and Digg stories [31]. This prediction strategy was motivated by a strong linear correlation observed between the (log-transformed) popularity of objects and earlier measures of user accesses (also log-transformed). For example, the authors observed Pearson linear correlations above 0.90 between the popularity of Digg stories measured at 1 hour and at 30 days after upload as well as between the popularity of YouTube videos measured at 7 and 30 days after upload. These correlations are stronger than those observed for tips. For example, the R^2 value of the regression from popularity in 1 week to popularity in 1 month is only 0.57 (for tips in G1) and 0.49 (for all tips), which correspond to linear correlations of 0.75 and 0.7, respectively[3]. For shorter monitoring periods t_r or longer values of δ, the R^2 values are much lower, indicating that popularity at time t_r can only explain a small fraction of the total popularity acquired by the tip at $t_r + \delta$.

Table 2: Rich-get-richer Analysis: coefficients α (and 95% confidence intervals) and R^2 of linear regressions from (log) popularity in t_r to (log) popularity $t_r + \delta$.

		Tips in G1		All tips	
$t_r+\delta$	t_r	α	R^2	α	R^2
1 mo	1 hr	0.799 ± 0.033	0.09	0.749 ± 0.011	0.07
1 mo	1 day	0.763 ± 0.016	0.26	0.822 ± 0.006	0.21
1 mo	1 wk	0.838 ± 0.009	0.57	0.887 ± 0.004	0.49
2 mo	1 day	0.594 ± 0.017	0.17	0.673 ± 0.007	0.13
2 mo	1 wk	0.681 ± 0.011	0.40	0.753 ± 0.004	0.31
2 mo	1 mo	0.834 ± 0.006	0.74	0.856 ± 0.003	0.65
6 mo	1 day	0.309 ± 0.015	0.07	0.397 ± 0.007	0.05
6 mo	1 wk	0.394 ± 0.010	0.20	0.489 ± 0.005	0.16
6 mo	1 mo	0.504 ± 0.008	0.40	0.562 ± 0.003	0.33

This result motivates the development of more sophisticated prediction models, such as those proposed in [35], which exploit other factors (e.g., characteristics of the user who posted the tip, the venue where it was posted and its content) to estimate the future popularity of a given tip. Yet, a different prediction task consists of estimating the ranking *by popularity* of a given *set* of tips at a future time. This is a possibly easier task, as it requires predicting not the actual popularity (or popularity level, as in [35]) of a tip but rather its relative popularity according to others. The prediction of popularity ranking supports various applications such as tip filtering and recommendation. Next, we evaluate the stability of tip popularity ranking over time, and assess to which extent the current popularity of a set of tips can be used to predict their future popularity ranking, and to which extent such prediction can be improved by also exploiting other features.

5. PREDICTING THE FUTURE

In the previous section, we observed that tips have longer lifespans than other types of online content (e.g, tweets, photos), and that tip popularity dynamics may be more strongly influenced by factors other than simply their current popularity (e.g., social network). We now further analyze this issue by assessing to which extent the *relative* popularity of a set of Foursquare tips can be predicted using only their popularity at prediction time, and to which extent the use of other attributes may improve prediction accuracy. In [35], we tackled the problem of predicting the popularity level of a given tip at posting time, formulating it as a classification task, and showed the importance of taking into account attributes of both the user who posted the tip and the venue where the tip was posted for that task.

[3]The R^2 is the square of the linear correlation between predictor and response variables.

We here focus on a different task, modeling the prediction as a *ranking task*, which aims at ranking a group of tips based on their predicted popularity at a future time. The ranking of the most popular tips helps to summarize a large set of tips focusing on the most popular ones for a scenario of interest (e.g., a city, a venue), instead of looking at the tips individually. By focusing on this task, we complement not only our prior prediction effort [35] but also our current analyses of tip popularity dynamics. Our ultimate goal is to assess to which extent the popularity ranking of a group of tips remains stable over time, and thus can be used to predict the ranking at a future time.

We first define our prediction task (Section 5.1), present the ranking strategies (Section 5.2) and the set of features (Section 5.3) used. We then discuss our experimental setup (Section 5.4) and results (Section 5.5).

5.1 Popularity Prediction Task

The problem we tackle can be formally defined as follows. Given a set P_d of tips posted in the previous d time units ($d \in (0, \infty)$) that meet a certain criterion c, rank those tips according to their expected popularity, measured in terms of the total number of likes they will receive up to time $t_r + \delta$, where t_r is the time when the ranking is performed. Criterion c may be, for example, tips posted at venues of a given city and/or category (e.g., Food), or even at a given venue. An empty criterion implies in no further constraint on set P_d.

Note that different tips in P_d may have been posted at different times within the time window $[t_r - d, t_r]$. Thus, we associate a posting time t_{p_i} with each tip p_i in P_d. We also consider sets V and U of venues and users, respectively, where $u_i \in U$ is the user who posted p_i, and $v_i \in V$ is the venue where it was posted. For evaluation purposes, we consider that each tip $p_i \in P_d$ is *labeled* with a numeric value that represents the number of likes received by p_i in the time interval $[t_{p_i}, t_r + \delta]$ (i.e., the true popularity acquired by p_i up to $t_r + \delta$), as discussed in Section 5.4. Each entity in P_d, U and V has a set of features F associated with it. Collectively, the features associated with p_i, u_i and v_i are used as inputs to a ranking model (see below) representing the given tip instance. The values of these features for a tip p_i are computed considering all the information available up to the time when the ranking is performed (t_r).

The choice of criterion c allows for different scenarios where the tip ranking problem becomes relevant. One scenario is that of a user who is interested in quickly finding tips with greater potential of becoming popular, and thus of containing valuable information, posted in any venue in her home city. A different scenario is that of a user who is particularly interested in retrieving tips regarding restaurants in her home city (or neighborhood). A business owner can also benefit from a ranking restricted to tips posted at venues of a specific category to get feedback about her business and about her competitors. Also, changes in the current and future tip popularity rankings can help with indirect analysis such as the influence of certain users whose tips got promoted in the future and the potential market share gains or losses for certain venues or venue categories.

5.2 Ranking Strategies

Recall that our goal is to assess to which extent using only the tips' current popularity ranking is enough to accurately predict their ranking at a future time. Thus, we here consider two ranking strategies. The first approach simply uses the ranking of the tips at prediction time (t_r) as an estimate of their ranking at the future time $t_r + \delta$. If the popularity ranking is stable, this approach should lead to perfect predictions. Thus, by analyzing the effectiveness of this approach we are indirectly assessing the stability of tip popularity ranking. We refer to this approach as *baseline*.

In order to assess the potential benefit of exploiting other factors to this prediction task, we consider a second approach that combines multiple features. To that end, we rely on an ordinary least square (OLS) multivariate regression model to predict the popularity of each tip p_i in P_d at time $t_r + \delta$ and then rank the tips by their predictions. In this approach, the logarithm of the number of likes of a tip p_i, $\mathcal{R}_t$, is estimated as a linear function of k predictor variables or features (presented in the next section), i.e.: $\mathcal{R}_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k$. Model parameters β_i (i=0..k) are determined by the minimization of the least squared errors [16] in the training data, as will be discussed in Section 5.4.

We note that various other algorithms could be used to exploit multiple features to predict the popularity ranking of a set of tips. Indeed, we did experiment with more sophisticated regression algorithms (notably Support Vector Regression (SVR) with radial basis function kernel [8], which handles non-linear relationships) as well as with a state-of-the-art learning-to-rank algorithm called Random Forests [5]. However, when applied with the same set of features, their results are similar (or even worse in some cases) than those obtained with the simpler OLS regression[4]. Thus, in order to avoid hurt readability, we present only OLS results.

5.3 Features

We consider a large set of features related to the three central entities which, intuitively, should be related to the popularity of a tip: textual content, user (i.e., tip's author), and venue. Specifically, we represent each tip p_i by $k = 53$ features related to the user u_i who posted p_i, the venue v_i where p_i was posted, and to the content of p_i. The values of these features are computed at the time when the ranking is performed (t_r). Table 3 shows the complete set of features. We have exploited most of these features for classifying a tip into low or high (predicted) popularity [35], although some features are novel and specific to the task of ranking multiple tips, as further discussed below. Some of these features, such as average number of likes received by all previous tips of user u_i and size of the tip p_i, have also been previously explored to analyze the helpfulness of online reviews [17, 38] and predict the ratings of (long) reviews [22, 29].

User features describe the tip's author past behavior and degree of activity in the system. Features related to the numbers of tips previously posted, number of likes received or given, and her social network are considered. Similarly, venue features capture the activity at the venue or its visibility to other users. For example, a tip may have a higher chance of becoming popular if it is posted at a venue that has more visibility. We also try to capture the strategy adopted by Foursquare to display the tips posted at the same venue,

[4]We note that we also found OLS to be as good as (if not better than) SVR when applied to the (different) task of predicting the popularity level of a given tip [35].

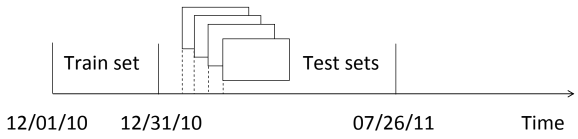


Figure 7: Temporal Data Split into Train and Test Sets

which may also impact the visibility of a tip by including the position of the tip in the rankings of tips of the venue.

We also consider features related to the tip's content. Numbers of characters and words, number of URLs or e-mails, as well as the fractions of words of each grammatical class are included. The latter are computed using the Stanford Part of Speech tagger[5], which employs probabilistic methods to build parse trees for sentences aiming at representing their grammatical structure, as in [21, 22]. We also include three features to represent sentiment scores obtained from SentiWordNet [9]. SentiWordNet is a lexical resource for supporting opinion mining by assigning three scores (positive, negative and neutral) to each synset (set of one or more synonyms) in the WordNet lexical database of English [11]. The scores are in the range of [0,1] and sum up to 1 for each word. We compute a positive, a negative and a neutral score for each tip by taking the average of the respective scores for each word in the tip that has an entry in SentiWordNet, as in [29]. To handle negation, we add the tag NOT to every word between a negation word (e.g., "no", "didn't") and the first punctuation mark following it [27], which implies that the positive scores of these words are converted to negative ones. Since some of our textual features are computed based on tools that were developed for English language only, we used a Linux dictionary (*myspell*) to filter tips with fewer than 60% of the words in English out from our datasets.

Since different tips in set P_d may have been posted at different times, we also add the age of the tip (in number of hours since posting time t_{p_i}) and the number of likes it has already received. These features are novel and have not been exploited in [35].

Tips can also be evaluated for their credibility as source of information. Fogg et al. [12] described credibility as a perceived quality composed by multiple dimensions, and showed that four website design elements – *Real-World Feel*, *Ease of Use*, *Expertise*, and *Trustworthiness* – impact credibility. Some of our features are based on these elements, as indicated in Table 3.

5.4 Experimental Setup

We build two scenarios to evaluate the prediction strategies: ranking all tips recently posted at venues located in New York, the city for which we have the largest number of tips, and ranking tips posted at venues of a specific category (Food) (also the largest category) located in New York [6]. In both scenarios, we consider only tips posted in the previous month (i.e., d = 30 days), and produce rankings based on their predicted popularity δ days later. We compare the effectiveness of both prediction strategies for various values of δ. Table 4 summarizes these two datasets, presenting the total numbers of tips, venues and users in each of them (the two rightmost columns are discussed below).

Unlike the baseline, the regression model needs to be parameterized. Thus, our experimental setup consists, in general terms, of dividing the data into training and test sets, using the former to learn the model parameters and the latter to evaluate the learned model. We split the tips chronologically into training and test sets, rather than randomly, to avoid including in the training tips that were posted after the tips for which predictions will be performed (test set). Figure 7 illustrates this chronological splitting used. For comparison purposes, we also evaluate the baseline only in the test sets.

The training set is composed of all tips posted from December 1^{st} to 30^{th}, 2010. These tips are used to learn the (regression-based) ranking model. We assume the ranking of the training instances is done on December 30^{th}, and thus use the total number of likes received by these tips at the target date (i.e., δ days later) as the ground truth to build the regression model.

Table 3: Features Used by the OLS Regression Model

Type	Description
User	Total # of tips posted by the user
	Number of of venues where the author posted tips
	Total # of likes received by previous tips of the author[1,c]
	Total # of likes given by the tip's author
	Number of friends or followers of the author
	Ratio of all likes received by the author coming from his friends and followers
	Total # of tips posted by the author's social network[1]
	# likes given by author's social network (in any tip)[1]
	Fraction of all likes received by the tip's author that are associated with tips posted at the same venue of the current tip but after it was posted[1]
	User category defined by Foursquare
	Total # of mayorships won by the author[a]
	If the author was mayor of the venue where tip was posted[a]
Venue	Total # of tips posted at the venue[b]
	Total # of likes received by tips posted at the venue[1,b]
	Total # of checkins at the venue[b]
	Total # of unique visitors[b]
	If the tipped venue was verified by Foursquare[c]
	Venue category defined by Foursquare
	Position of the tip in the tips of the venue sorted by # of likes in ascending order
	Position of the tip in the ranking of the venue sorted by # of likes in descending order
	Position of the tip in the ranking of the venue sorted by date in ascending order
Content	# of likes received until time t_r
	Hours since posting until time t_r
	Length of the text of the tip, in characters
	Length of the text of the tip, in number of words
	# of URLs or emails address contained on a tip
	Fraction of nouns in the tip
	Fraction of adjectives in the tip
	Fraction of adverbs in the tip
	Fraction of comparatives in the tip
	Fraction of verbs in the tip
	Fraction of non-English words in the tip
	Fraction of numbers in the tip
	Fraction of superlatives in the tip
	Fraction of symbols in the tip
	Fraction of punctuation in the tip
	Average positive score over all words in the tip
	Average neutral score over all words in the tip
	Average negative score over all words in the tip

[1] Median, average and standard deviation are also included.
[a] Based on Fogg's design element Expertise.
[b] Based on the Fogg's design element Trustworthiness.
[c] Based on the Fogg's design element Real-world feel.

[5]www-nlp.stanford.edu/software/corenlp.shtml
[6]Other scenarios, such as ranking tips posted at a venue, are also possible. However, the highly skewed distribution of tips per venue leads to severe data sparsity, which, in turn, poses a challenge to the training of the regression model.

Table 4: Overview of Datasets and Scenarios of Evaluation

Scenarios	# of tips	# of users	# of venues	# of tips in training sets	Avg # of tips in test sets
NY	169,393	55,149	31,737	516	4,697.87
NY Food	81,742	32,961	8,927	244	2,365.0

Recall that the distribution of number of likes per tip is highly skewed towards very few number of likes (Section 3), which might bias the regression model and ultimately hurt its accuracy[7]. Thus, we adopt the following approach to reduce this skew. We group tips in the *training set* according to a threshold τ for the number of likes received by the tip at the target date. Two classes are defined: all tips with at least τ likes are grouped into the high popularity class, and the others are grouped into the low popularity class. We then build balanced training sets according to these two classes by performing under-sampling: we randomly select n tips from the low popularity class, whereas n is the number of tips in the high popularity class[8]. We repeat this process r times, thus building multiple (balanced) training sets. We experiment with various values of τ finding best results with τ=5. This was also the threshold used in [35] to predict the *popularity level* of a tip at a future target date. However, whereas in that work the classification task was the core of the prediction strategy, here it is employed simply for evaluation purposes (i.e., for balancing the training set). We also use r=5 replications, which allows us to assess the variability of our results. We note that this under-sampling approach (and threshold τ) is applied *only to the training set.* The test sets (described next) remains unchanged (imbalanced). Table 4 (5^{th} column) presents the total number of tips in the training sets for each scenario.

We then use tips posted from December 31^{st} until February 27^{th} 2011 to build 30 different test sets, as follows. Since tips can be continually liked, the predicted ranking may become stale. Thus, we evaluate the effectiveness of the ranking methods by using them to build a new ranking by the end of each day (starting on January 29^{th}), always considering the tips posted in the previous $d = 30$ days. Thus, 30 test sets are built by considering a window of 30 days and sliding it 1 day at a time, 30 times. Table 4 (6^{th} column) shows the average number of tips in each test set[9]. For each test set, we report average results produced by all 5 training sets, and corresponding 95% confidence intervals.

For both training and test sets, the features of each tip are computed using all data collected up to the time when ranking is performed (t_r), including (for the regression model) information associated with tips posted before the beginning of each training set. Moreover, feature values are computed by first applying a logarithm transformation on the raw numbers to reduce their large variability, and then scaling these results between 0 and 1. We note that, in order to have enough historical data about users who posted tips, for both training and test sets, we consider only tips posted by users with at least five tips. We determine the best parameters of the regression models by minimizing the least squared errors of predictions for the candidate tips in the training set.

We evaluate each ranking method by computing the Kendall τ rank distance of the top-k tips in the rankings produced by it (i.e., Kτ@k). Since we are comparing two top-k lists (τ_1 and τ_2), we use a modified Kendall τ metric [18], that uses a penalty parameter p, with $0 \leq p \leq 1$, to account for the distances between non-overlapping tips in τ_1 and τ_2[10]. The modified Kendall τ is defined as follows:

$$K\tau(\tau_1,\tau_2)@k = (k - |\tau_1 \cap \tau_2|)((2+p)k - p|\tau_1 \cap \tau_2| + 1 - p) + \sum_{i \in \tau_i \cap \tau_2} \kappa_{i,j}(\tau_1,\tau_1) - \sum_{i \in \tau_1 - \tau_2} \tau_1(i) - \sum_{i \in \tau_2 - \tau_1} \tau_2(i) \quad (1)$$

where $\tau_1(i)$ or $\tau_2(i)$ is the position in the rank of the ith item and $\kappa_{i,j}(\tau_1,\tau_2) = 0$ if $\tau_1(i) < \tau_1(j)$ and $\tau_2(i) < \tau_2(j)$, or $\kappa_{i,j}(\tau_1,\tau_2) = 1$, otherwise. K$\tau$@k ranges from 0 to 1, with values close to 1 indicating greater disagreement between the predicted ranking and an ideal ranking defined by the actual number of likes accumulated by each tip until t_r+δ (i.e., the tip's label).

5.5 Experimental Results

We discuss our results by first assessing how the popularity ranking of tips varies over time (Section 5.5.1), and then comparing the prediction based only on the current ranking (baseline) and the regression-based prediction that uses a richer set of features (Section 5.5.2).

5.5.1 Ranking Stability

Using the experimental setup described in Section 5.4, we investigate the differences between the *true* popularity rankings of tips at times t_r and $t_r + \delta$, for various values of δ. To that end, we quantify the correlation between these two rankings using Kendall's τ coefficient. Recall that the closer to 1 the value of Kτ is, the larger the disagreements between both rankings.

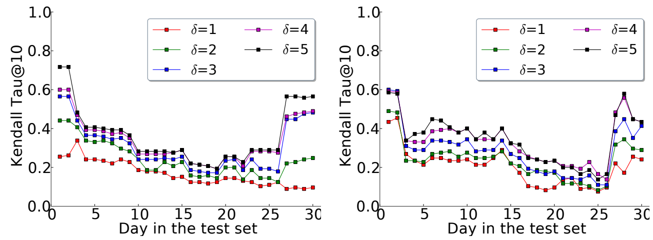
Figure 8 shows the Kτ@k for each day in the test fold of both NY and NY Food scenarios, for values of δ varying from 1 to 5 months. We focus on the top-10 most popular tips (k=10). Focusing first on the NY scenario, Figure 8a shows that the disagreements between both rankings increase as we increase δ. Indeed, for a fixed test day (fixed set of tips), the Kτ@10 varies from 0.26 to 0.72 as we increase δ from 1 to 5 months. Moreover, we can still observe some discrepancies even if we predict for only one month ahead in the future (δ=1 month). Indeed, as discussed in Section 4, over 40% of the likes of most tips arrive after two months since posting time. Since the tips in each test fold are at most 1 month old, most of them are still at very early stages of their popularity curves, and the popularity ranking, even considering only the top-10 tips, will change. Very similar results were also observed for the NY Food scenario, as shown in Figure

[7]Great imbalance in the training set, as observed in our datasets, is known to have a detrimental impact on the effectiveness of classification and regression algorithms.

[8]For illustration purposes, we note that the original training set for the NY scenario had 5,225 tips in the low popularity class and only 258 tips in the other (smaller) class.

[9]The results are qualitatively similar when ranking is performed at lower frequencies, once each k days ($k > 1$).

[10]We use $p = 0.5$ which was recommended by [10].



(a) NY (b) NY Food

Figure 8: Correlations between the top-10 most popular tips at time t_r and at time $t_r + \delta$ (δ in months).

8b, although the values of Kτ@10 (and thus the disagreements between current and future rankings) seem somewhat smaller on some days, particularly for larger values of δ.

Examining the top most popular tips in each test fold for the NY scenario, we found that some of them referred to special events occurring in the city. These tips exhibit a somewhat different pattern: all of their likes are received until the event occurs. Thus, once they reach the top of the ranking, they tend to remain there for a while, which contributes to lower the discrepancies between predicted and future rankings.

Overall, these results corroborate our discussion in Section 4, and suggest that there are some noticeable discrepancies between the current and the long-term popularity of tips (even within the top-10 most popular tips). Thus, models that use only early measurements may lead to inaccurate predictions not only of popularity measures (as discussed in Section 4.4) but also of popularity ranking. Next, we assess to which extent such ranking predictions can be improved by exploiting a multidimensional set of predictors.

5.5.2 *Prediction Results*

We now compare the prediction results using only the popularity ranking at t_r (baseline) against the prediction produced by using the OLS regression model jointly with the features defined in Section 5.3. Figure 9 shows the average daily Kτ@10 along with 95% confidence intervals for the two ranking methods and each value of δ, for the NY scenario. For δ equal to 1 month, both methods produce τ@k results below 0.4, showing a high correlation between the predicted ranking and the true popularity ranking at $t_r + \delta$. However, the OLS regression model produces results that are significantly better (lower Kτ@10) than those produced by the baseline in 67% of the days (reductions in up to 69%).

Moreover, as we predict further into the future, increasing δ to 2, 5 and 6 months, we observe increasing values of Kτ@10 for both methods. This implies that the discrepancies with the true ranking tend to increase as both methods start using outdated and possibly inaccurate data. Yet, the gap between the baseline and the OLS regression model tends to increase (reaching up to 65% for δ equal to 6 months). This result shows that taking factors other than simply the current popularity of the tips into account is important and can improve prediction accuracy of the long-term popularity ranking.

We note, however, that there are some cases where the baseline performs as good as the more sophisticated OLS model. These specific cases are explained by the following: some of the most popular tips (which referred to real events), acquired most of their likes very early on before time t_r (before the event). Thus, they quickly reached top positions of the ranking, remaining there until $t_r+\delta$. For such cases, the use of other features produces only marginal improvements in prediction.

Figure 10 shows similar results for the NY Food scenario. In this case, we see smaller differences between both methods. In most cases, the baseline is just as good as the more sophisticated OLS method, although the use of the extra features does provide improvements (up to 30%) in some of the days for large values of δ. These results reflect the higher stability of the tip popularity ranking in the NY Food scenario (Section 5.5.1). Moreover, as shown in Table 4, the number of tips in the training set of this scenario is almost half of that used in the NY scenario, which also impacts the accuracy of the regression model. That is, the benefits from using more features are constrained by the limited amount of data to train an accurate model[11]

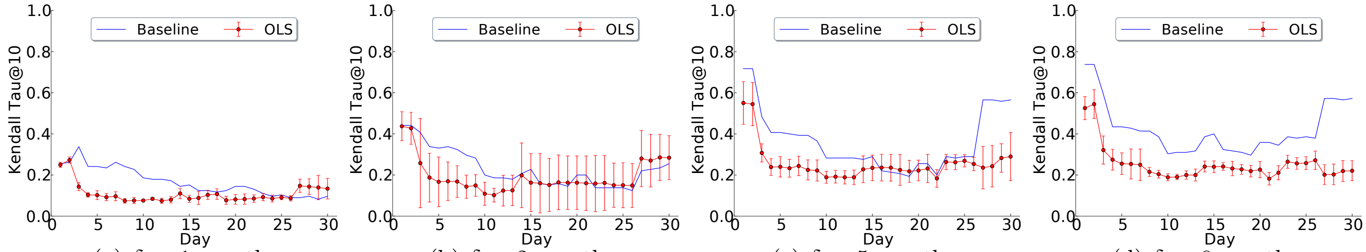
These results highlight that the accurate popularity prediction of tips is a challenging task. Although tip popularity ranking remains roughly stable over short periods of time (e.g., 1 month), there are still significant discrepancies that occur in the top of the ranking. Moreover, the use of other features related to the tip's author, venue and tip's content can improve prediction accuracy to some extent, provided that enough information about the features is available to train the model.

Finally, we sorted the features used by the OLS method using the Information Gain feature selection technique [37]. We found that the most important feature is, unsurprisingly, the current popularity of the tip. It is followed by features related to the user's popularity, such as the total number of likes in previous tips. Features related to the social network of the tip's author (number of followers and friends, and average number of tips posted by them) are also in the top-10 most important features, consistently with our results in Section 4.2.

The most important venue feature is the total number of check-ins, followed by the current position of the tip in the ranking of tips of the venue sorted by increasing number of likes. However, these features, like the other venue related features, are much less important than the user features, occupying only the 24^{th} and 25^{th} positions of the ranking. Similarly, the most important content feature is the number of characters in the tip, but it occupies only the 21^{st} position of the ranking. Thus, like observed in [35] and unlike in other efforts to assess the helpfulness of online reviews [17, 38], textual features play a much less important role in the tip popularity ranking prediction task, possibly due to the inherent different nature of these pieces of content.

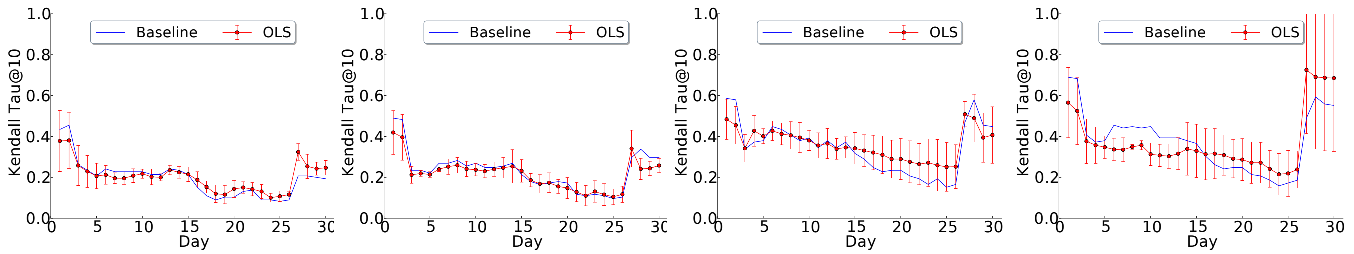
We did test whether multicollinearity exists among different predictors, which could affect robustness of the results of the OLS model. In our analysis, we use two methods: variance inflation factors (VIF) [30] and tolerance [25]. For each predictor variable j, $VIF_j = \frac{1}{1-R_j^2}$, where R_i^2 is the coefficient of determination from a regression using predictor j as response variable and all the other predictors as independent variables. A VIF value greater than 10 is a indication

[11] Recall that we did experiment with other prediction strategies based on SVR and Random Forests, but OLS provided the best results across all scenarios.



(a) δ = 1 month (b) δ = 2 months (c) δ = 5 months (d) δ = 6 months

Figure 9: Effectiveness of Ranking for Varying Target Time $t_r+\delta$: NY Scenario (average and 95% confidence intervals)



(a) δ = 1 month (b) δ = 2 months (c) δ = 5 months (d) δ = 6 months

Figure 10: Effectiveness of Ranking for Varying Target Time $t_r+\delta$: NY Food Scenario (average and 95% confidence intervals)

of potential multicollinearity problem [30]. The tolerance is the reciprocal of VIF, and is computed as $1 - VIF$. The smaller the tolerance value (< 0.10), the higher the degree of collinearity [25]. We compute the VIFs and tolerances for all features of our OLS model and we found that despite the strong correlations between some pairs of predictors, removing some of these variables from the model does not have impact on the accuracy (Kτ@10).

6. CONCLUSION AND FUTURE WORK

We presented an extensive study of tip popularity dynamics using a large dataset collected from Foursquare. Although prior work has tackled the popularity dynamics of various types of user generated content, we are not aware of prior temporal analysis of online reviews. We found that most tips have a slow popularity evolution, acquiring most of their likes after a few months, and that the social network of the tip's author plays an important role to draw attention to the tip, particularly soon after posting time. We also found that most tips reach their daily popularity peak within a month in the system, although most of their likes are received after the peak. Moreover, compared to other types of content, we observed a weaker presence of the rich-get-richer phenomenon, indicating a lower correlation between the early and long-term popularities of the tip. This suggests that the tip popularity prediction may require more sophisticated models, exploring other factors related to the tip, besides their current popularity.

We further analyzed this issue by looking into the stability of the popularity ranking over time, observing that there are some noticeable disagreements between the current and future popularity rankings, even when considering only the top-10 most popular tips and a time window of only 1 month. This suggests that predicting the future ranking based only on the current ranking may not be accurate. We thus investigated to which extent we can improve such predictions by using a regression model and exploiting a multidimensional set of features related to the tip's author, the venue where it was posted and its content. Our results show that the use of these features can improve the prediction accuracy, given that enough training data is available.

As future work, we intend to analyze the temporal dynamics of user tipping and "liking" activities, and their correlations with tip popularity, and investigate the benefits for prediction of adding new features, particularly geographic related features, such as the distance between different venues where a user posts tips.

7. ACKNOWLEDGEMENTS

This research is partially funded by the Brazilian National Institute of Science and Technology for Web Research (MCT/CNPq/ INCT grant number 573871/2008-6), CNPq, CAPES and Fapemig.

8. REFERENCES

[1] R. Bandari, S. Asur, and B. Huberman. The Pulse of News in Social Media: Forecasting Popularity. In *Proc. of ICWSM*, 2012.

[2] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] A. Bermingham and A. Smeaton. Classifying Sentiment in Microblogs: is Brevity an Advantage? In *Proc. of CIKM*, 2010.

[4] Y. Borghol, S. Ardon, N. Carlsson, D. L. Eager, and A. Mahanti. The Untold Story of the Clones: Content-Agnostic Factors that Impact YouTube Video Popularity. In *Proc. of SIGKDD*, 2012.

[5] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001.

[6] M. Cha, A. Mislove, and K. P. Gummadi. A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In *Proc. of WWW*, 2009.

[7] R. Crane and D. Sornette. Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System. In *Proc. of PNAS*, volume 105, pages 15649–15653, 2008.

[8] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vladimir. Support Vector Regression Machines. In *Proc. of NIPS*, 1997.

[9] A. Esuli and F. Sebastiani. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *In Proc. of LREC*, 2006.

[10] R. Fagin, R. Kumar, and D. Sivakumar. Comparing Top K Lists. In *Proc. of SODA*, 2003.

[11] C. Fellbaum. *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA, 1998.

[12] B. Fogg, J. Marshall, O. Laraki, A. Osipovich, et al. What Makes Web Sites Credible?: a Report on a Large Quantitative Study. In *Proc. of CHI*, 2001.

[13] A. Ghose and P. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE TKDE*, 23(10):1498–1512, 2011.

[14] L. Hong, O. Dan, and B. D. Davison. Predicting Popular Messages in Twitter. In *Proc. of WWW*, 2011.

[15] Y. Hong, J. Lu, J. Yao, Q. Zhu, and G. Zhou. What Reviews are Satisfactory: Novel Features for Automatic Helpfulness Voting. In *Proc. of SIGIR*, 2012.

[16] R. Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley, 1991.

[17] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically Assessing Review Helpfulness. In *Proc. of EMNLP*, 2006.

[18] A. Konagurthu and J. Collier. An Information Measure for Comparing Top k Lists. *CoRR*, abs/1310.0110, 2013.

[19] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of Growing Random Networks. *Physical Review Letters*, 85(21):4629–4632, Nov. 2000.

[20] K. Lerman and R. Ghosh. Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. In *ICWSM*, 2010.

[21] Y. Liu, X. Huang, A. An, and X. Yu. Modeling and Predicting The Helpfulness of Online Reviews. In *Proc. of the ICDM*, 2008.

[22] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting Social Context for Review Quality Prediction. In *Proc. of WWW*, 2010.

[23] Y. Matsubara, Y. Sakurai, A. Prakash, L. Li, and C. Faloutsos. Rise and Fall Patterns of Information Diffusion: Model and Implications. In *Proc. of the KDD*, 2012.

[24] E. Momeni, C. Cardie, and M. Ott. Properties, Prediction, and Prevalence of Useful User-Generated Comments for Descriptive Annotation of Social Media Objects. In *Proc. of ICWSM*, 2013.

[25] R. O'Brien. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity: International Journal of Methodology*, 41(5):673–690, October 2007.

[26] M. O'Mahony and B. Smyth. Learning to Recommend Helpful Hotel Reviews. In *Proc. of RecSys*, 2009.

[27] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In *Proc. of EMNLP*, 2002.

[28] H. Pinto, J. Almeida, and M. Gonçalves. Using Early View Patterns to Predict the Popularity of YouTube Videos. In *Proc. of WSDM*, 2013.

[29] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro. How Useful are Your Comments?: Analyzing and Predicting YouTube Comments and Comment Ratings. In *Proc. of WWW*, 2010.

[30] J. Stevens. *Applied Multivariate Statistics for The Social Sciences*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2002.

[31] G. Szabo and B. A. Huberman. Predicting the Popularity of Online Content. *Communications of the ACM*, 53(8):80–88, Aug. 2010.

[32] A. Tatar, P. Antoniadis, M. Amorim, and S. Fdida. From Popularity Prediction to Ranking Online News. *Soc. Netw. Anal. and Min.*, page 4:174, Jan. 2014.

[33] T. Valente. *Network Models of the Diffusion of Inovations*. Hampton Press, Cresskill, NJ, 1995.

[34] R. van Zwol. Flickr: Who is Looking? In *Web Intelligence*, 2007.

[35] M. Vasconcelos, J. Almeida, and G. Marcos. What Makes your Opinion Popular? Predicting the Popularity of Micro-Reviews in Foursquare. In *Proc. of SAC*, 2014.

[36] J. Yang and J. Leskovec. Patterns of Temporal Variation in Online Media. In *Proc. of*, 2009.

[37] Y. Yang and J. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proc. of ICML*, 1997.

[38] Z. Zhang and B. Varadarajan. Utility Scoring of Product Reviews. In *Proc. of CIKM*, 2006.