Everyone's an Influencer: Quantifying Influence on Twitter

Eytan Bakshy*
University of Michigan, USA
ebakshy@umich.edu

Jake M. Hofman
Yahoo! Research, NY, USA
hofman@yahoo-inc.com

Winter A. Mason
Yahoo! Research, NY, USA
winteram@yahoo-inc.com

Duncan J. Watts
Yahoo! Research, NY, USA
djw@yahoo-inc.com

ABSTRACT

In this paper we investigate the attributes and relative influence of 1.6M Twitter users by tracking 74 million diffusion events that took place on the Twitter follower graph over a two month interval in 2009. Unsurprisingly, we find that the largest cascades tend to be generated by users who have been influential in the past and who have a large number of followers. We also find that URLs that were rated more interesting and/or elicited more positive feelings by workers on Mechanical Turk were more likely to spread. In spite of these intuitive results, however, we find that predictions of which particular user or URL will generate large cascades are relatively unreliable. We conclude, therefore, that word-of-mouth diffusion can only be harnessed reliably by targeting large numbers of potential influencers, thereby capturing average effects. Finally, we consider a family of hypothetical marketing strategies, defined by the relative cost of identifying versus compensating potential "influencers." We find that although under some circumstances, the most influential users are also the most cost-effective, under a wide range of plausible assumptions the most cost-effective performance can be realized using "ordinary influencers"—individuals who exert average or even less-than-average influence.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems; J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms

Human Factors

*Part of this research was performed while the author was visiting Yahoo! Research, New York.

WSDM'11, February 9–12, 2011, Hong Kong, China.

Keywords

Communication networks, Twitter, diffusion, influence, word of mouth marketing.

1. INTRODUCTION

Word-of-mouth diffusion has long been regarded as an important mechanism by which information can reach large populations, possibly influencing public opinion [14], adoption of innovations [26], new product market share [4], or brand awareness [15]. In recent years, interest among researchers and marketers alike has increasingly focused on whether or not diffusion can be maximized by seeding a piece of information or a new product with certain special individuals, often called "influentials" [34, 15] or simply "influencers," who exhibit some combination of desirable attributes—whether personal attributes like credibility, expertise, or enthusiasm, or network attributes such as connectivity or centrality—that allows them to influence a disproportionately large number of others [10], possibly indirectly via a cascade of influence [31, 16].

Although appealing, the claim that word-of-mouth diffusion is driven disproportionately by a small number of key influencers necessarily makes certain assumptions about the underlying influence process that are not based directly on empirical evidence. Empirical studies of diffusion are therefore highly desirable, but historically have suffered from two major difficulties. First, the network over which word-of-mouth influence spreads is generally unobservable, hence influence is difficult to attribute accurately, especially in instances where diffusion propagates for multiple steps [29, 21]. And second, observational data on diffusion are heavily biased towards "successful" diffusion events, which by virtue of being large are easily noticed and recorded; thus inferences regarding the attributes of success may also be biased [5, 9], especially when such events are rare [8].

For both of these reasons, the micro-blogging service Twitter presents a promising natural laboratory for the study of diffusion processes. Unlike other user-declared networks (e.g. Facebook), Twitter is expressly devoted to disseminating information, in that users subscribe to broadcasts of other users; thus the network of "who listens to whom" can be reconstructed by crawling the corresponding "follower graph". In addition, because users frequently wish to share web-content, and because tweets are restricted to 140 characters in length, a popular strategy has been to use URL

shorteners (e.g. bit.ly, TinyURL, etc.), which effectively tag distinct pieces of content with unique, easily identifiable tokens. Together these features allow us to track the diffusion patterns of all instances in which shortened URLs are shared on Twitter, regardless of their success, thereby addressing both the observability and sampling difficulties outlined above.

The Twitter ecosystem is also well suited to studying the role of influencers. In general, influencers are loosely defined as individuals who disproportionately impact the spread of information or some related behavior of interest [34, 10, 15, 11]. Unfortunately, however, this definition is fraught with ambiguity regarding the nature of the influence in question, and hence the type of individual who might be considered special. Ordinary individuals communicating with their friends, for example, may be considered influencers, but so may subject matter experts, journalists, and other semi-public figures, as may highly visible public figures like media representatives, celebrities, and government officials. Clearly these types of individuals are capable of influencing very different numbers of people, but may also exert quite different types of influence on them, and even transmit influence through different media. For example, a celebrity endorsing a product on television or in a magazine advertisement presumably exerts a different sort of influence than a trusted friend endorsing the same product in person, who in turn exerts a different sort of influence than a noted expert writing a review.

In light of this definitional ambiguity, an especially useful feature of Twitter is that it not only encompasses various types of entities, but also forces them all to communicate in roughly the same way: via tweets to their followers. Although it remains the case that even users with the same number of followers do not necessarily exert the same kind of influence, it is at least possible to measure and compare the influence of individuals in a standard way, by the activity that is observable on Twitter itself. In this way, we avoid the need to label individuals as either influencers or non-influencers, simply including all individuals in our study and comparing their impact directly.

We note, however, that our use of the term influencer corresponds to a particular and somewhat narrow definition of influence, specifically the user's ability to post URLs which diffuse through the Twitter follower graph. We restrict our study to users who "seed" content, meaning they post URLs that they themselves have not received through the follower graph. We quantify the influence of a given post by the number of users who subsequently repost the URL, meaning that they can be traced back to the originating user through the follower graph. We then fit a model that predicts influence using an individual's attributes and past activity and examine the utility of such a model for targeting users. Our emphasis on prediction is particularly relevant to our motivating question. In marketing, for example, the practical utility of identifying influencers depends entirely on one's ability to do so in advance. Yet in practice, it is very often the case that influencers are identified only in retrospect, usually in the aftermath of some outcome of interest, such as the unexpected success of a previously unknown author or the sudden revival of a languishing brand [10]. By emphasizing ex-ante prediction of influencers over ex-post explanation, our analysis highlights some simple but useable insights that we believe are of general relevance to word-of-mouth marketing and related activities.

The remainder of the paper is organized as follows. We review related work on modeling diffusion and quantifying influence in Section 2. In Sections 3 and 4 we provide an overview of the collected data, summarizing the structure of URL cascades on the Twitter follower graph. In Section 5, we present a predictive model of influence, in which cascade sizes of posted URLs are predicted using the individuals' attributes and average size of past cascades. Section 6 explores the relationship between content as characterized by workers on Amazon's Mechanical Turk and cascade size. Finally, in Section 7 we use our predictive model of cascade size to examine the cost-effectiveness of targeting individuals to seed content.

2. RELATED WORK

A number of recent empirical papers have addressed the matter of diffusion on networks in general, and the attributes and roles of influencers specifically. In early work, Gruhl et al [13] attempted to infer a transmission network between bloggers, given time-stamped observations of posts and assuming that transmission was governed by an independent cascade model. Contemporaneously, Adar and Adamic [1] used a similar approach to reconstruct diffusion trees among bloggers, and shortly afterwards Leskovec et al. [20] used referrals on an e-commerce site to infer how individuals are influenced as a function of how many of their contacts have recommended a product.

A limitation of these early studies was the lack of "ground truth" data regarding the network over which the diffusion was taking place. Addressing this problem, more recent studies have gathered data both on the diffusion process and the corresponding network. For example, Sun et al. [29] studied diffusion trees of fan pages on Facebook, Bakshy et al. [3] studied the diffusion of "gestures" between friends in Second Life, and Aral et al. [2] studied adoption of a mobile phone application over the Yahoo! messenger network. Most closely related to the current research is a series of recent papers that examine influence and diffusion on Twitter specifically. Kwak et al. [18] compared three different measures of influence—number of followers, page-rank, and number of retweets—finding that the ranking of the most influential users differed depending on the measure. Cha et al. [7] also compared three different measures of influence—number of followers, number of retweets, and number of mentions—and also found that the most followed users did not necessarily score highest on the other measures. Finally, Weng et al. [35] compared number of followers and page rank with a modified page-rank measure that accounted for topic, again finding that ranking depended on the influence measure.

The present work builds on these earlier contributions in three key respects. First, whereas previous studies have quantified influence either in terms of network metrics (e.g. page rank) or the number of direct, explicit retweets, we measure influence in terms of the size of the entire diffusion tree associated with each event (Kwak et al [18] also compute what they call "retweet trees" but they do not use them as a measure of influence). While related to other measures, the size of the diffusion tree is more directly associated with diffusion and the dissemination of information (Goyal et al [12], it should be noted, do introduce a similar metric to quantify influence; however, their interest is in identifying community

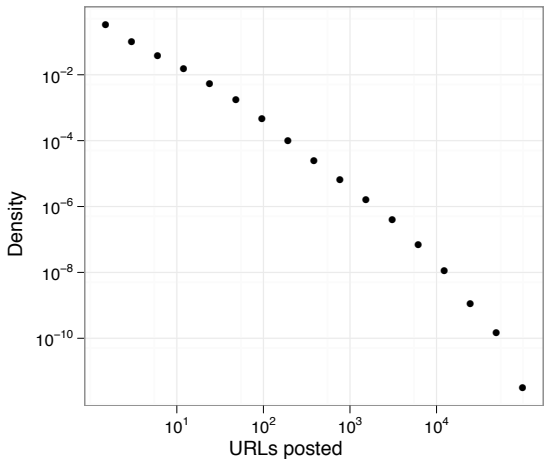


Figure 1: Probability density of number of bit.ly URLs posted per user

"leaders," not on prediction.) Second, whereas the focus of previous studies has been largely descriptive (e.g. comparing the most influential users), we are interested explicitly in predicting influence; thus we consider all users, not merely the most influential. Third, in addition to predicting diffusion as a function of the attributes of individual seeds, we also study the effects of content. We believe these differences bring the understanding of diffusion on Twitter closer to practical applications, although as we describe later, experimental studies are still required.

3. DATA

To study diffusion on Twitter, we combined two separate but related sources of data. First, over the two-month period of September 13 2009 - November 15 2009 we recorded all 1.03B public tweets broadcast on Twitter, excluding October 14-16 during which there were intermittent outages in the Twitter API. Of these, we extracted 87M tweets that included bit.ly URLs and which corresponded to distinct diffusion "events," where each event comprised a single initiator, or "seed," followed by some number of repostings of the same URL by the seed's followers, their followers, and so on[1]. Finally, we identified a subset of 74M diffusion events that were initiated by seed users who were active in both the first and second months of the observation period; thus enabling us to train our regression model on first month performance in order to predict second-month performance (see Section 5). In total, we identified 1.6M seed users who seeded an average of 46.33 bit.ly URLs each. Figure 1 shows the distribution of bit.ly URL posts by seed.

Second, we crawled the portion of the follower graph comprising all users who had broadcast at least one URL over the same two-month period. We did this by querying the Twitter API to find the followers of every user who posted a bit.ly URL. Subsequently, we placed those followers in a queue to be crawled, thereby identifying their followers, who were then also placed in the queue, and so on. In this way, we obtained a large fraction of the Twitter follower graph comprising all active bit.ly posters and anyone connected to these users via one-way directed chains of followers. Specifically, the subgraph comprised approximately 56M users and 1.7B edges.

Consistent with previous work [7, 18, 35], both the in-degree ('followers") and out-degree ("friends") distributions are highly skewed, but the former much more so—whereas the maximum # of followers was nearly 4M, the maximum # of friends was only about 760K—reflecting the passive and one-way nature of the "follow" action on Twitter (i.e. A can follow B without any action required from B). We emphasize, moreover, that because the crawled graph was seeded exclusively with active users, it is almost certainly not representative of the entire follower graph. In particular, active users are likely to have more followers than average, in which case we would expect that the average in-degree will exceed the average out-degree for our sample—as indeed we observe. Table 1 presents some basic statistics of the distributions of the number of friends, followers and number of URLs posted per user.

Table 1: Statistics of the Twitter follower graph and seed activity

	# Followers	# Friends	# Seeds Posted
Median	85.00	82.00	11.00
Mean	557.10	294.10	46.33
Max.	3,984,000.00	759,700.00	54,890

4. COMPUTING INFLUENCE ON TWITTER

To calculate the influence score for a given URL post, we tracked the diffusion of the URL from its origin at a particular "seed" node through a series of reposts—by that user's followers, those users' followers, and so on—until the diffusion event, or cascade, terminated. To do this, we used the time each URL was posted: if person B is following person A, and person A posted the URL before B and was the only of B's friends to post the URL, we say person A influenced person B to post the URL. As Figure 2 shows, if B has more than one friend who has previously posted the same URL, we have three choices for how to assign the corresponding influence: first, we can assign full credit to the friend who posted it first; second we can assign full credit to the friend who posted it most recently (i.e. last); and third, we can split credit equally among all prior-posting friends.

These three assignments effectively make different assumptions about the influence process: "first influence" rewards primacy, assuming that individuals are influenced when they first see a new piece of information, even if they fail to immediately act on it, during which time they may see it again; "last influence" assumes the opposite, instead attributing influence to the most recent exposure; and "split influence" assumes either that the likelihood of noticing a new piece of information, or equivalently the inclination to act on it, accumulates steadily as the information is posted by more

[1]Our decision to restrict attention to bit.ly URLs was made predominantly for convenience, as bit.ly was at the time by far the dominant URL shortener on Twitter. Given the size of the population of users who rely on bit.ly, which is comparable to the size of all active Twitter users, it seems unlikely to differ systematically from users who rely on other shorteners.

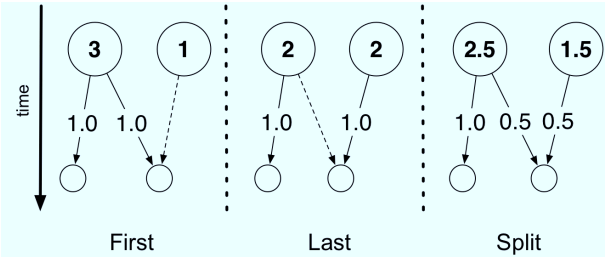


Figure 2: Three ways of assigning influence to multiple sources

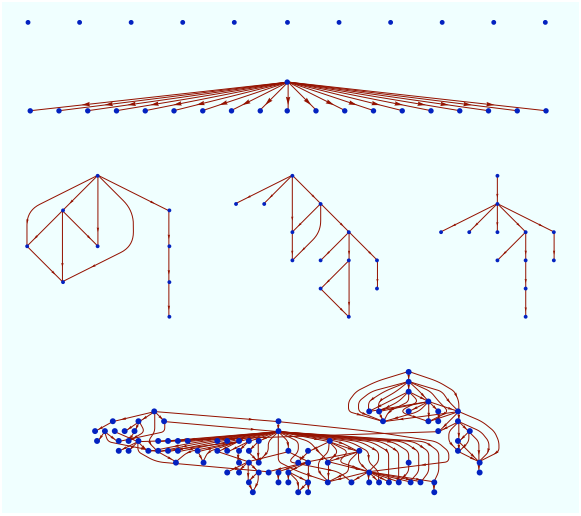


Figure 3: Examples of information cascades on Twitter.

friends. Having defined immediate influence, we can then construct disjoint influence trees for every initial posting of a URL. The number of users in these influence trees—referred to as "cascades"—thus define the influence score for every seed. See Figure 3 for some examples of cascades. To check that our results are not an artifact of any particular assumption about how individuals are influenced to repost information, we conducted our analysis for all three definitions. Although particular numerical values varied slightly across the three definitions, the qualitative findings were identical; thus for simplicity we report results only for first influence.

Before proceeding, we note that our use of reposting to indicate influence is somewhat more inclusive than the convention of "retweeting" (e.g. using the terminology "RT @username") which explicitly attributes the original user. An advantage of our approach is that we can include in our observations all instances in which a URL was reposted regardless of whether it was acknowledged by the user, thereby greatly increasing the coverage of our observations. (Since our study, Twitter has introduced a "retweet" feature that arguably increases the likelihood that reposts will be acknowledged, but does not guarantee that they will be.) However, a potential disadvantage of our definition is that it may mistakenly attribute influence to what is in reality a sequence of independent events. In particular, it is likely that users who follow each other will have similar interests and so are more likely to post the same URL in close succession than random pairs of users. Thus it is possible that some of what we are labeling influence is really a consequence of homophily [2]. From this perspective, our estimates of influence should be viewed as an upper bound.

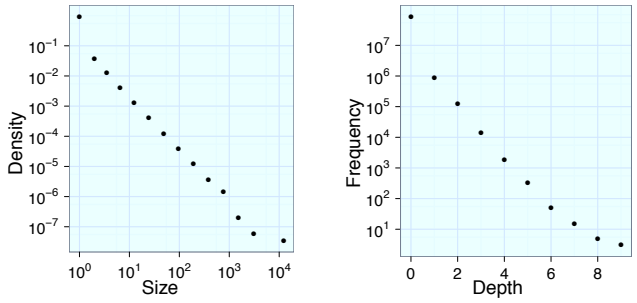
On the other hand, there are reasons to think that our measure underestimates actual influence, as re-broadcasting a URL is a particularly strong signal of interest. A weaker but still relevant measure might be to observe whether a given user views the content of a shortened URL, implying that they are sufficiently interested in what the poster has to say that they will take some action to investigate it. Unfortunately click-through data on bit.ly URLs are often difficult to interpret, as one cannot distinguish between programmatic unshortening events—e.g., from crawlers or browser extensions—and actual user clicks. Thus we instead relied on reposting as a conservative measure of influence, acknowledging that alternative measures of influence should also be studied as the platform matures.

Finally, we reiterate that the type of influence we study here is of a rather narrow kind: being influenced to pass along a particular piece of information. As we discuss later, there are many reasons why individuals may choose to pass along information other than the number and identity of the individuals from whom they received it—in particular, the nature of the content itself. Moreover, influencing another individual to pass along a piece of information does not necessarily imply any other kind of influence, such as influencing their purchasing behavior, or political opinion. Our use of the term "influencer" should therefore be interpreted as applying only very narrowly to the ability to consistently seed cascades that spread further than others. Nevertheless, differences in this ability, such as they do exist, can be considered a certain type of influence, especially when the same information (in this case the same original URL) is seeded by many different individuals. Moreover, the terms "influentials" and "influencers" have often been used in precisely this manner [3]; thus our usage is also consistent with previous work.

5. PREDICTING INDIVIDUAL INFLUENCE

We now investigate an idealized version of how a marketer might identify influencers to seed a word-of-mouth campaign [16], where we note that from a marketer's perspective the critical capability is to identify attributes of individuals that consistently predict influence. Reiterating that by "influence" we mean a user's ability to seed content containing URLs that generate large cascades of reposts, we therefore begin by describing the cascades we are trying to predict.

As Figure 4a shows, the distribution of cascade sizes is approximately power-law, implying that the vast majority of posted URLs do not spread at all (the average cascade size is 1.14 and the median is 1), while a small fraction are reposted thousands of times. The depth of the cascade (Figure 4b) is also right skewed, but more closely resembles an exponential distribution, where the deepest cascades can propagate as far as nine generations from their origin; but again the vast majority of URLs are not reposted at all, corresponding to cascades of size 1 and depth 0 in which the seed is the only node in the tree. Regardless of whether



(a) *Cascade Sizes* (b) *Cascade Depths*

Figure 4: (a). Frequency distribution of cascade sizes. (b). Distribution of cascade depths.

we study size or depth, therefore, the implication is that most events do not spread at all, and even moderately sized cascades are extremely rare.

To identify consistently influential individuals, we aggregated all URL posts by user and computed individual-level influence as the logarithm of the average size of all cascades for which that user was a seed. We then fit a regression tree model [6], in which a greedy optimization process recursively partitions the feature space, resulting in a piecewise-constant function where the value in each partition is fit to the mean of the corresponding training data. An important advantage of regression trees over ordinary linear regression (OLR) in this context is that unlike OLR, which tends to fit the vast majority of small cascades at the expense of larger ones, the piecewise constant nature of the regression tree function allows cascades of different sizes to be fit independently. The result is that the regression tree model is much better calibrated than the equivalent OLR model. Moreover, we used folded cross-validation [25] to terminate partitioning to prevent over-fitting. Our model included the following features as predictors:

1. Seed user attributes
   (a) # followers
   (b) # friends
   (c) # tweets,
   (d) date of joining
2. Past influence of seed users
   (a) average, minimum, and maximum total influence
   (b) average, minimum, and maximum local influence,

where past local influence refers to the average number of reposts by that user's immediate followers in the first month of the observation period, and past total influence refers to average total cascade size over the same period. Followers, friends, number of tweets, and influence (actual and past) were all log-transformed to account for their skewed distributions. We then compared predicted influence with actual influence computed from the second month of observations.

Figure 5 shows the regression tree for one of the folds. Conditions at the nodes indicate partitions of the features, where the left (right) child is followed if the condition is satisfied (violated). Leaf nodes give the predicted influence—as measured by (log) mean cascade size— for the corresponding partition. Thus, for example, the right-most leaf indicates that users with upwards of 1870 followers who had on average 6.2 reposts by direct followers (past local influence) are predicted to have the largest average total influence, generating cascades of approximately 8.7 additional posts.

Unsurprisingly, the model indicates that past performance provides the most informative set of features, although it is the local, not the total influence that is most informative; this is likely due to the fact that most non-trivial cascades are of depth 1, so that past direct adoption is a reliable predictor of total adoption. Also unsurprisingly, the number of followers is an informative feature. Notably, however, these are the only two features present in the regression tree, enabling us to visualize influence as a function of these features, as shown in Figure 6. This result is somewhat surprising, as one might reasonably have expected that individuals who follow many others, or very few others, would be distinct from the average user. Likewise, one might have expected that activity level, quantified by the number of tweets, would also be predictive.

Figure 7 shows the fit of the regression tree model for all five cross-validation folds. The location of the circles indicates the mean predicted and actual values at each leaf of the trees, with leaves from different cross-validation folds appearing close to each other; the size of the circles indicates the number of points in each leaf, while the bars show the standard deviation of the actual values at each leaf. The model is extremely well calibrated, in the sense that the prediction of the average value at each cut of the regression tree is almost exactly the actual average ($R^2 = 0.98$). This appearance, however, is deceiving. In fact, the model fit without averaging predicted and actual values at the leaf nodes is relatively poor ($R^2 = 0.34$), reflecting that although large cascades tend to be driven by previously successful individuals with many followers, the extreme scarcity of such cascades means that most individuals with these attributes are not successful either. Thus, while large follower count and past success are likely necessary features for future success, they are far from sufficient.

These results place the usual intuition about influencers in perspective: individuals who have been influential in the past and who have many followers are indeed more likely to be influential in the future; however, this intuition is correct only on average. We also emphasize that these results are based on far more observational data than is typically available to marketers—in particular, we have an objective measure of influence and extensive data on past performance. Our finding that individual-level predictions of influence nevertheless remain relatively unreliable therefore strongly suggests that rather than attempting to identify exceptional individuals, marketers seeking to exploit word-of-mouth influence should instead adopt portfolio-style strategies, which target many potential influencers at once and therefore rely only on average performance [33].

6. THE ROLE OF CONTENT

An obvious objection to the above analysis is that it fails to account for the nature of the content that is being shared. Clearly one might expect that some types of content (e.g. YouTube videos) might exhibit a greater tendency to spread

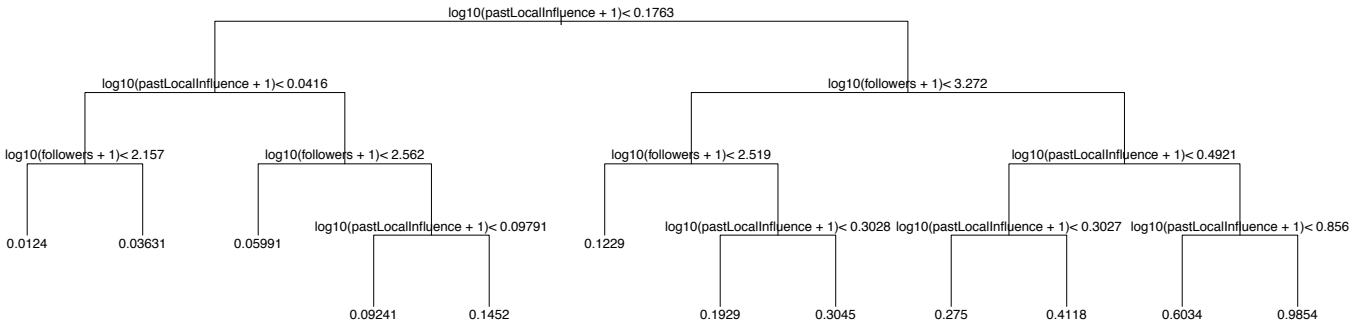
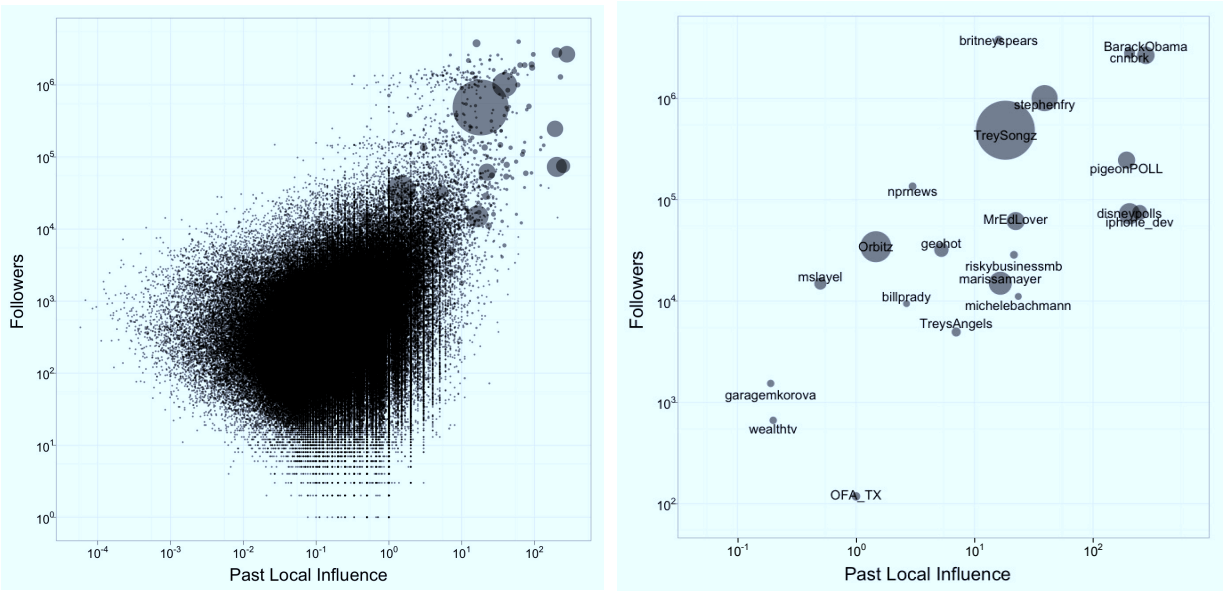


Figure 5: Regression tree fit for one of the five cross-validation folds. Leaf nodes give the predicted influence for the corresponding partition, where the left (right) child is followed if the node condition is satisfied (violated).



(a) *All users* (b) *Top 25 users*

Figure 6: Influence as a function of past local influence and number of followers for (a) all users and (b) users with the top 25 actual influence. Each circle represents a single seed user, where the size of the circle represents that user's actual average influence.

than others (e.g. news articles of specialized interest), or that even the same type of content might vary considerably in interestingness or suitability for sharing. Conceivably, one could do considerably better at predicting cascade size if in addition to knowing the attributes of the seed user, one also knew something about the content of the URL being seeded.

To test this idea, we used humans to classify the content of a sample of 1000 URLs from our study. An advantage of this approach over an automated classifier or a topic-model [35] is that humans can more easily rate content on attributes like "interestingness" or "positive feeling" which are often quite difficult for a machine. A downside of using humans, however, is that the number of URLs we can classify in this way is necessarily small relative to the total sample. Moreover, because the distribution of cascade sizes is so skewed, a uniform random sample of 1000 URLs would almost certainly not contain any large cascades; thus we instead obtained a stratified sample in the following manner. First, we filtered URLs that we knew to be spam or in a language other than English. Second, we binned all remaining URLs in logarithmic bins choosing an exponent such that we obtained ten bins in total, and the top bin contained the 100 largest cascades. Third, we sampled all 100 URLs in the top bin, and randomly sampled 100 URLs from each of the remaining bins. In this way, we ensured that our sample would reflect the full distribution of cascade sizes.

Given this sample of URLs, we then used Amazon's Mechanical Turk (AMT) to recruit human classifiers. AMT is a system that allows one to recruit workers to do small tasks for small payments, and has been used extensively for survey and experimental research [17, 24, 23, 22] and to obtain labels for data [28, 27]. We asked the workers to go to the web page associated with the URL and answer questions about it. Specifically, we asked them to classify the site as "Spam / Not Spam / Unsure", as "Media Sharing / Social Networking, Blog / Forum, News / Mass Media, or

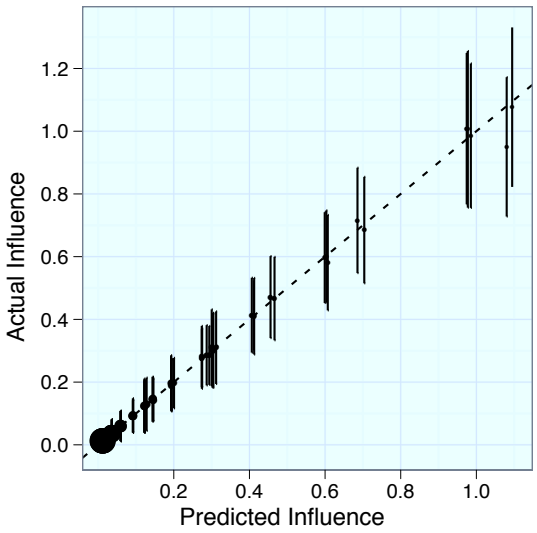


Figure 7: Actual vs. predicted influence for regression tree. The model assigns each seed user to a leaf in the regression tree. Points representing the average actual influence values are placed at the prediction made by each leaf, with vertical lines representing one standard deviation above and below. The size of the point is proportional to the number of diffusion events assigned to the leaf. Note that all five folds are represented here, including the fold represented in Figure 5, where proximate lines correspond to leaves from different cross-validation folds.

Other", and then to specify one of 10 categories it fell into (see Figure 8b). We then asked them to rate how broadly relevant the site was, on a scale from 0 (very niche) to 100 (extremely broad). We also gauged their impression of the site, including how interesting they felt the site was (7-point Likert scale), how interesting the average person would find it (7-point Likert scale), and how positively it made them feel (7-point Likert scale). Finally, we asked them to indicate if they would share the URL using any of the following services: Email, IM, Twitter, Facebook, or Digg.

To ensure that our ratings and classifications were reliable, we had each URL rated at least 3 times—the average URL was rated 11 times, and the maximum was rated 20 times. If more than three workers marked the URL as a bad link or in a foreign language, it was excluded. In addition, we excluded URLs that were marked as spam by the majority of workers. This resulted in 795 URLs that we could analyze.

As Figure 9 shows, content that is rated more interesting tends to generate larger cascades on average, as does content that elicits more positive feelings. In addition, Figure 8 shows that certain types of URLs, like those associated with shareable media, tend to spread more than URLs associated with news sites, while some types of content (e.g "lifestyle") spread more than others.

To evaluate the additional predictive power of content, we repeat the regression tree analysis of Section 5 for this subset of URLs, adding the following content-based features:

1. Rated interestingness
2. Perceived interestingness to an average person
3. Rated positive feeling
4. Willingness to share via Email, IM, Twitter, Facebook or Digg.
5. Indicator variables for type of URL (see Figure 8a)
6. Indicator variable for category of content (see Figure 8b)

Figure 10 shows the model fit including content. Surprisingly, none of the content features were informative relative to the seed user features (hence we omit the regression tree itself, which is essentially identical to Figure 5), nor was the model fit ($R^2 = 0.31$) improved by the addition of the content features. We note that the slight decrease in fit and calibration compared to the content-free model can be attributed to two main factors: first, the training set size is orders of magnitude smaller for the content model as we have fewer hand-labeled URLs, and second, here we are making predictions at the single post level, which has higher variance than the user-averaged influence predicted in the content-free model.

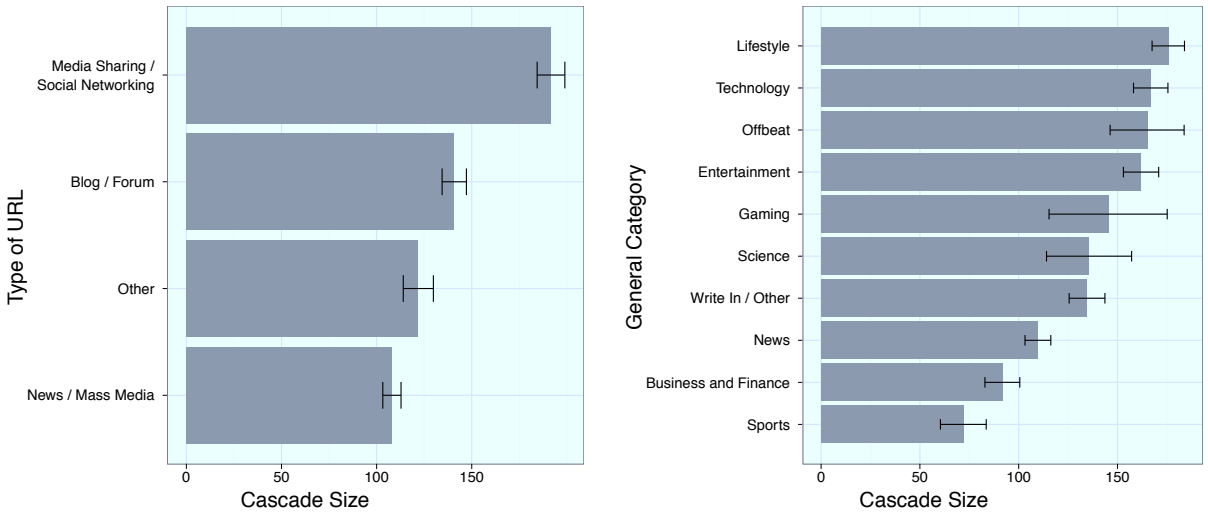
These results are initially surprising, as explanations of success often do invoke the attributes of content to account for what is successful. However, the reason is essentially the same as above—namely that most explanations of success tend to focus only on observed successes, which invariably represent a small and biased sample of the total population of events. When the much larger number of non-successes are also included, it becomes difficult to identify content-based attributes that are consistently able to differentiate success from failure at the level of individual events.

7. TARGETING STRATEGIES

Although content was not found to improve predictive performance, it remains the case that individual-level attributes—in particular past local influence and number of followers—can be used to predict average future influence. Given this observation, a natural next question is how a hypothetical marketer might exploit available information to optimize the diffusion of information by systematically targeting certain classes of individuals. In order to answer such a question, however, one must make some assumptions regarding the costs of targeting individuals and soliciting their cooperation.

To illustrate this point we now evaluate the cost-effectiveness of a hypothetical targeting strategy based on a simple but plausible family of cost functions $c_i = c_a + f_i c_f$, where c_a represents a fixed "acquisition cost" c_a per individual i, and c_f represents a "cost per follower" that each individual charges the marketer for each "sponsored" tweet. Without loss of generality we have assumed a value of $c_f = \$0.01$, where the choice of units is based on recent news reports of paid tweets (http://nyti.ms/atfmzx). For convenience we express the acquisition cost as multiplier α of the per-follower cost; hence $c_a = \alpha c_f$.

Because the relative cost of targeting potential "influencers" is an unresolved empirical question, we instead explore a



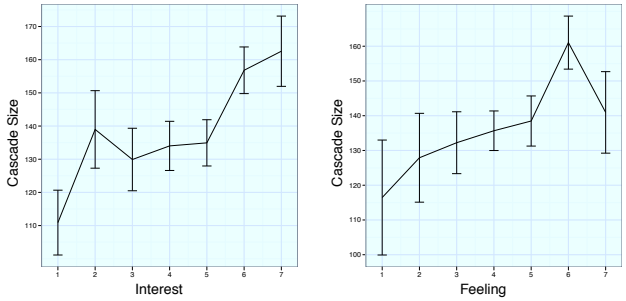
(a) *Type of URL* (b) *Content Category for URL*

Figure 8: (a). Average cascade size for different type os URLs (b). Average cascade size for different categories of content. Error bars are standard errors

wide range of possible assumptions by varying α. For example, choosing α to be small corresponds to a cost function that is biased towards individuals with relatively few followers, who are cheap and numerous. Conversely when α is sufficiently large, the acquisition cost will tilt toward targeting a small number of highly influential users, meaning users with a larger number of followers and good track records. Regardless, one must trade off between the number of followers per influencer and the number of individuals who can be targeted, where the optimal tradeoff will depend on α. To explore the full range of possibilities allowed by this family of cost functions, for each value of α we binned users according to their influence as predicted by the regression tree model and computed the average influence-per-dollar of the targeted subset for each bin.

As Figure 11 shows, when $\alpha = 0$—corresponding to a situation in which individuals can be located costlessly—we find that by far the most cost-effective category is to target the least influential individuals, who exert over fifteen times the influence-per-dollar of the most influential category. Although these individuals are much less influential (average influence score ≈ 0.01) than average, they also have relatively few followers (average ≈ 14); thus are relatively inexpensive. At the other extreme, when α become sufficiently large—here $\alpha \gtrsim 100,000$, corresponding to an acquisition cost $c_a = \$1,000$—we recover the result that highly influential individuals are also the most cost-effective. Although expensive, these users will be preferred simply because the acquisition cost prohibits identifying and managing large numbers of influencers.

Finally, Figure 11 reveals that although the most cost-efficient category of influencers corresponds to increasingly influential individuals as α increases, the transition is surprisingly slow. For example, even for values of α as high as 10,000, (i.e. equivalent to $c_a = \$100$) the most cost-efficient influencers are still relatively ordinary users, who exhibit approximately average influence and connectivity.



(a) *Interesting* (b) *Positive Feeling*

Figure 9: (a). Average cascade size for different interest ratings (b). Average cascade size for different ratings of positive feeling. Error bars are standard errors.

8. CONCLUSIONS

In light of the emphasis placed on prominent individuals as optimal vehicles for disseminating information [19], the possibility that "ordinary influencers"—individuals who exert average, or even less-than-average influence—are under many circumstances more cost-effective, is intriguing. We emphasize, however, that these results are based on statistical modeling of observational data and do not imply causality. It is quite possible, for example, that content seeded by outside sources—e.g., marketers—may diffuse quite differently than content selected by users themselves. Likewise, while we have considered a wide range of possible cost functions, other assumptions about costs are certainly possible and may lead to different conclusions. For reasons such as these, our conclusions therefore ought to be viewed as hypotheses to be tested in properly designed experiments, not as verified causal statements. Nevertheless, our find-

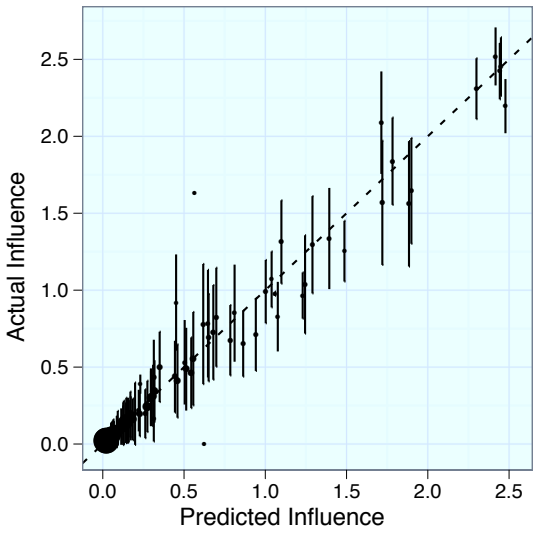


Figure 10: Actual vs. predicted influence for regression tree including content

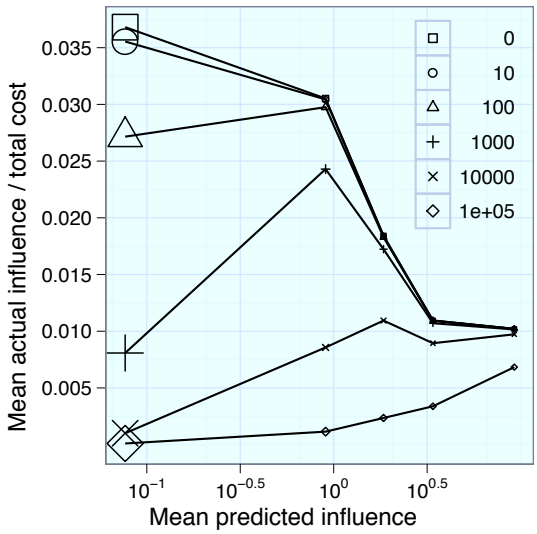


Figure 11: Return on investment where targeting cost is defined as $c_i = c_a + f_i c_f$, $c_a = \alpha c_f$ and $\alpha \in \{0, 10, 100, 1000, 10000, 100000\}$.

ing regarding the relative efficacy of ordinary influencers is consistent with previous theoretical work [32] that has also questioned the feasibility of word-of-mouth strategies that depend on triggering "social epidemics" by targeting special individuals.

We also note that although Twitter is in many respects a special case, our observation that large cascades are rare is likely to apply in other contexts as well. Correspondingly, our conclusion that word-of-mouth information spreads via many small cascades, mostly triggered by ordinary individuals, is also likely to apply generally, as has been suggested elsewhere [33]. Marketers, planners and other change agents interested in harnessing word-of-mouth influence could therefore benefit first by adopting more precise metrics of influence; second by collecting more and better data about potential influencers over extended intervals of time; and third, by potentially exploiting ordinary influencers, where the optimal tradeoff between the number of individuals targeted and their average level of influence will depend on the specifics of the cost function in question.

9. ACKNOWLEDGMENTS

We thank Sharad Goel for helpful conversations.

10. REFERENCES

[1] E. Adar and A. Adamic, Lada. Tracking information epidemics in blogspace. In *2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Compiegne University of Technology, France, 2005.

[2] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544, 2009.

[3] E. Bakshy, B. Karrer, and A. Adamic, Lada. Social influence and the diffusion of user-created content. In *10th ACM Conference on Electronic Commerce*, Stanford, California, 2009. Association of Computing Machinery.

[4] F. M. Bass. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, 1969.

[5] R. A. Berk. An introduction to sample selection bias in sociological data. *American Sociological Review*, 48(3):386–398, 1983.

[6] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. Chapman & Hall/CRC, 1984.

[7] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummad. Measuring user influence on twitter: The million follower fallacy. In *4th Int'l AAAI Conference on Weblogs and Social Media*, Washington, DC, 2010.

[8] R. M. Dawes. *Everyday irrationality: How pseudo-scientists, lunatics, and the rest of us systematically fail to think rationally*. Westview Pr, 2002.

[9] J. Denrell. Vicarious learning, undersampling of failure, and the myths of management. *Organization Science*, 14(3):227–243, 2003.

[10] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Little Brown, New York, 2000.

[11] J. Goldenberg, S. Han, D. R. Lehmann, and J. W. Hong. The role of hubs in the adoption process. *Journal of Marketing*, 73(2):1–13, 2009.

[12] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. Discovering leaders from community actions. pages 499–508. ACM, 2008. Proceeding of the 17th ACM conference on Information and knowledge management.

[13] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. pages 491–501. ACM New York, NY, USA, 2004.

[14] E. Katz and P. F. Lazarsfeld. *Personal influence; the part played by people in the flow of mass communications*. Free Press, Glencoe, Ill.,, 1955.

[15] E. Keller and J. Berry. *The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy*. Free Press, New York, NY, 2003.

[16] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA., 2003. Association of Computing Machinery.

[17] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 2008.

[18] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? pages 591–600. ACM, 2010.

[19] A. Leavitt, E. Burchard, D. Fisher, and S. Gilbert. The influentials: New approaches for analyzing influence on twitter.

[20] J. Leskovec, A. Adamic, Lada, and A. Huberman, Bernardo. The dynamics of viral marketing. *ACM Trans. Web*, 1(1):5, 2007.

[21] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633, 2008.

[22] W. Mason and S. Suri. Conducting Behavioral Research on Amazon's Mechanical Turk. *SSRN eLibrary*, 2010.

[23] W. Mason and D. J. Watts. Financial incentives and the performance of crowds. *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 77–85, 2009.

[24] S. A. Munson and P. Resnick. Presenting diverse political opinions: how and how much. pages 1457–1466. ACM, 2010.

[25] R. R. Picard and R. D. Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984.

[26] E. M. Rogers. *Diffusion of innovations*. Free Press, New York, 4th edition, 1995.

[27] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. pages 614–622. ACM, 2008.

[28] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast-but is it good? evaluating non-expert annotations for natural language tasks, 2008.

[29] E. S. Sun, I. Rosenn, C. A. Marlow, and T. M. Lento. Gesundheit! modeling contagion through facebook news feed. In *International Conference on Weblogs and Social Media*, San Jose, CA, 2009. AAAI.

[30] B. Tomlinson and C. Cockram. Sars: Experience at prince of wales hospital, hong kong. *The Lancet*, 361(9368):1486–1487, 2003.

[31] D. J. Watts. A simple model of information cascades on random networks. *Proceedings of the National Academy of Science, U.S.A.*, 99:5766–5771, 2002.

[32] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34:441–458, 2007.

[33] D. J. Watts and J. Peretti. Viral marketing for the real world. *Harvard Business Review*, May:22–23, 2007.

[34] G. Weimann. *The Influentials: People Who Influence People*. State University of New York Press, Albany, NY, 1994.

[35] J. Weng, E. P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. pages 261–270. ACM, 2010.