CS224W Final Project Report
Analyzing and Predicting Pinterest Influencers

David Hershey, Elliot Lui, Mathieu Rolfo

December 11, 2016

Abstract

In this project, we examine various predictors of influence in a social network. First, we define several forms of 'influence,' which correspond to a user's impact on other users' behavior over time. We then propose predictor metrics to predict the influence of nodes, and deploy algorithms that locate these users given a static snapshot of a social network. We consider a wide range of characteristics that could predict a node's influence, from classic graph statistics like centrality to more social network specific characteristics such as 'node diversity.' We perform analysis on data obtained from Pinterest, a content-sharing social network. We then compare the effectiveness of our predictor metrics in accurately predicting influencers. We find that PageRank predicts best when evaluating a single node, while number of pins is best amongst a larger sample. Overall, we find that PageRank, Most Pins, and Diversity are better than Degree Centrality and Most Boards at predicting node influence.

1 Introduction

As social networks have grown rapidly in recent years, there has been an increased interest in identifying the most influential actors, or the 'influencers,' in a network. This interest comes both from companies and brands interested in marketing products on social networks[2], as well as social scientists interested in understanding the patterns of behavior on and social structure of these networks[1][3][9]. Loosely defined, an influencer is someone who disproportionately impacts the spread of some quantity (usually information) across the network. However, the meaning of the term 'influencer' can vary by context and social network, depending on the objective of the task: someone could be interested in getting clicks on a specific piece of content, or obtaining more permanent "followers" to be able to get more exposure in the future. On networks like Twitter, one can calculate the number of likes and retweets, while on other networks different behaviors must be captured. Additionally, influence is a function of time: usually one wants to extrapolate from past actions and responses whether or not a future action will garner a large response. We examine a variety of metrics capturing this, building off the existing literature and proposing some additional metrics.

Given these components of influence, the social network Pinterest is an interesting and relatively unexplored domain for study, in comparison to other social networks. Pinterest is a content-sharing site that allows users to 'pin' images from around the web to various 'boards' that they create. The site has over 100 million active users and is one of the fastest-growing social networks. Previous work on Pinterest has largely examined its demographic trends[4][5]: the site is unusual for having primarily female users. In Gilbert et al.'s research, they examine whether location and gender predict user activity, finding statistically significant results for both of these properties[5]. We had access to temporal data about the network, with ordered lists of edges created over time, rather than the graph snapshots most other researchers have been using.

This paper seeks to determine metrics that are good predictors of influencers in our dataset - we call these 'predictor metrics'. Other studies have examined metrics like degree centrality and found them to be reasonable predictors of influence. There are other less common metrics we wish to evaluate. Liu et al. have defined a metric of diversity capturing how dissimilar that node's neighbors are from one another[6]. However, given the variety of metrics used in the literature, with results that contradict commonsense reasoning[3], we seek to empirically test previous findings on different networks and introduce new results for other researchers to verify.

2 Related Work

Research on other social media have taken a similar methodological approach to predicting and defining influence. In particular, Bakshy et al. and Cha et al. quantify predictive metrics on Twitter and define influence outcome metrics. We adapted their notion of URL diffusion throughout follower networks on Twitter as influence to repins and follows on Pinterest. Also, we sought to determine whether Cha et al.'s finding that indegree did not correlate with other measures of influence (i.e. the million follower fallacy) would be upheld on Pinterest. Other approaches to this problem have used similarity models to quantify the effectiveness of a leader [9]. Lappas et al. develop the notion of effectors in trying to understand which nodes are most likely to have created the observed activation state, a different metric than discussed in our paper but nonetheless a definition of influence. They find that prolific nodes are often effectors, which lends credence to our intuition that nodes with many boards and pins will be influencers in our Pinterest graph. Shafiq et al. explore influence by defining similarity metrics between the content produced by 'leaders' vs the content produced by 'followers'. This metric does quantify how much the leaders effect the ideas in the network, but is less applicable for following the spread of single pieces of content or for actions like 'follows' in a network. Future work on Pinterest might examine the content of pins and use this as a predictor of influence.

Regarding Pinterest specifically, we found both academic and industry writings on the site and user behavior. While non-technical, Cario's book on Pinterest Marketing shows the growing interest in understanding the site's dynamics and motivates the goal of finding influencers on the site. On the other hand, Chang et al.'s and Gilbert and Bakhshi's papers reflect the academic literature analyzing Pinterest. Chang et al. among other topics examine how similarity between users drives repinning on the site, motivating our investigations into the diversity metrics used in this project. Gilbert and Bakhshi provide evidence that number of pins and boards also predict social influence (repins), which we attempt to replicate here.

We also draw upon algorithmic literature in our paper. Aside from the metrics and approaches discussed above, we reference two other algorithms. One is the classic PageRank from Brin and Page, which we use as a predictor metric due to its success in identifying influential and important web pages. We also adapt the method introduced by *Liu et al* in Mining Diversity called Diversity that characterize's how diverse a node's connection to its neighborhood is. As a metric with a specific qualitative purpose, it offers an interesting angle as a predictor metric. *Liu et al* applies this metric to a few genres of networks including a synthetic network and a social network, but the data from the social network Renren used was relatively small with only 5,000 users and didn't have the temporal dimension to it.

3 Approach and Methods

The main goal of this work is to locate *influencers* in a social network: nodes that are the most influential on the behavior of their *followers* (defined as the set of nodes that have directed edges to a leader node). As direct calculation of influence can be computationally very expensive, the goal of this work is to predict influencer nodes in a static snapshot of a social network on the basis of other characteristics, which will correlate highly with actual influence scores. First, we discuss our predictor metrics and reasons for selecting them. Next, we discuss our quantifications of influence that we use as the gold standard for our predicted influencers.

3.1 Predictor Metrics

Here, we outline our predictor metrics to identify predicted influencers and our justifications for these metrics. We utilize four traditional metrics, some of which are Pinterest-specific, and one less-common metric that captures membership in various communities.

3.1.1 Degree Centrality

The most straightforward metric to identify influential nodes is degree centrality. We look for nodes with both high in-degree (many followers). The hypothesis is that a user that has a lot of followers and is very active in the network will likely have high influence on its followers. For this project, we use degree centrality to refer to in-degree, but filtering for a threshold value of

out-degree greater than 20. We do this because of how we calculate influence; if a node does not follow many users, then it is difficult to calculate their follower uptake.

3.1.2 PageRank

One important metric of node influence is the PageRank metric first employed by *Page et al*[7]. This metric seeks to identify the nodes in the network with the most and highest quality followers. This metric gives extra weight to followers that are also highly influential. As such, we expect that PageRank will identify nodes that are both highly effective and influential, based on the algorithm's success in the search domain.

3.1.3 Pins and Boards

We also examine metrics specific to Pinterest. We calculated the total number of boards created by a user, and the total number of pins pinned on a user's boards. Other researchers have successfully built models that use number of boards and number of pins to predict repins[5], leading us to believe it may be effective here. Additionally, we expect this metric to be effective due to the relationship between activity and influence: users that are most active on the site may be most noticed by others, and so become influential.

3.1.4 Diversity

Another more complex metric that could quantify a leader is the notion of a diverse node. The diversity of a node is defined as how different the node's connections to its neighborhood is [5]. We hypothesize that a node that has followers from different perspectives has a higher chance of garnering followers of different backgrounds. We see here in Figure 1 a few examples of diversity.

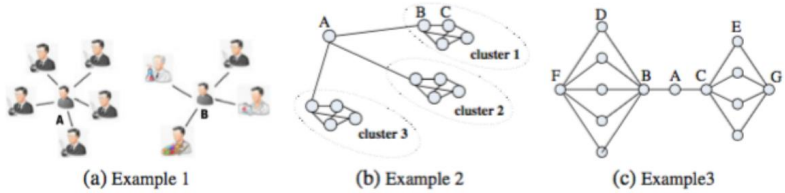

Figure 1

In Example 2, while node A and node C have the same degree, it is intuitive that node A has much higher diversity since each of its neighbors are part of three separate communities. We implement the algorithm discussed in *Liu lu et al* to find these high diversity nodes[6].

To calculate diversity for a single node n, the algorithm looks at n's r-neighborhood, where r is the radius of the neighborhood, and calculates how similar (or close) this set is to each other. So $r = 1$ means looking at just immediate neighbors to see if they are connected to each other, and an $r = 2$ looks at neighbors of neighbors. The running time of this algorithm ran in O(NM) time, where N is the number of nodes and M is the average r-neighborhood per node. We see as we look at larger values of r, if r is equal to the diameter of the network, then this becomes O(N^2) time. Because of this limitation, we ran this algorithm with $r = 2$.

3.2 Influence Metrics

In order to quantify the influence of a node, we develop a series of metrics that capture the intuitive notion of influence: the extent to which a node's followers mimic its behavior over time. High scores on these metrics show that the behavior of an influencer drives the evolution of the network.

3.2.1 Uptake Probability

The first metric we define is the follower *uptake probability*. This metric is defined as the average probability that when a node l follows another node n at time t_1, one of l's followers at time t_1 then follows n at time $t_2 > t_1$. Formalized

$$UP = \sum_{n \in S} \frac{1}{|S|} P(f \rightarrow_{t_2} n | l \rightarrow_{t_1} n) \quad (1)$$

Where $a \rightarrow_t b$ is defined as the event that a follows b at time t, f is a follower of l, S is the set of l's followers, and $t_2 > t_1$. The most influential possible node would have a uptake probability of 1.0, where all of that node's actions are copied by some followers.

3.2.2 Uptake Count

We also define the follower *uptake count*, defined as the average number of followers of node l that follow n after l follows n. Formalized:

$$UC = \frac{1}{|S|} \sum_{n \in S} \sum_{f \in F} \mathbb{1}(f \rightarrow_{t_2} n | l \rightarrow_{t_1} n) \quad (2)$$

Where F is the set of followers of l at time t_1. This metric quantifies the gross impact of a node on the network, so it scales more with popular nodes than with highly effective nodes.

The two follow influence metrics have significant value to brands, who may want to identify which nodes in a network would direct traffic to the brand's board if they are followed by influencers.

3.2.3 Average Follower Re-Pins

To quantify another type of influence, we also look at the spread of the content generated by a node throughout its local network. Specifically, we quantify the average number of re-pins a post made by a node gets from its set of followers. This metric has obvious value to advertisers as well, as they may want to identify whom to pay to seed a post in order to get maximum visibility in a network and convert to purchases.

4 Pinterest Food Dataset

We examined a subset of the Pinterest network corresponding to food-related topics. The full dataset has 10,389,475 users, 12,466,754 boards, and 20,049,957 unique pins. However, due to relative sparsity of the data, we did not have demographic information (e.g. gender, age, or location) for users. We did have access to the time at which boards were created and followed, and when pins were added to boards.

The dataset was initially structured as three separate graphs: follows of boards (user to board), pins to boards (pin to board), and creation of boards (user to board). From this data, we synthesized a user-user follower network, where a directed edge from user A to user B represents that A follows one of B's boards. As B can have multiple boards, this graph allows for multi-edges. Aside from pin and board metrics, all graph statistics were calculated on this synthesized graph.

As displayed by Figure 3, the graph is clearly a power law distributed network, with very similar distributions for in and out degree.

5 Analysis and Results

We identified our predicted influencers on a snapshot of the network at the time of the 1,000,000th follow. This gives us an initial set of users to evaluate the time-dependent metrics upon. We then calculated the influence metrics for these predicted influencers' nodes by letting the network evolve from the time the first follow occurred to the time the 20,000,000th follow occurred (over one year of real data). This allowed us to evaluate which predictor metrics were able to find the most influential nodes in the network while remaining computationally feasible.

In order to compare methods of identifying influencers, we used each predictor algorithm outlined above to identify the top 20 predicted influencers in each category on the 1,000,000 follow subset of the network: degree centrality, PageRank, number of boards and pins, and diversity. Due to the extreme computational cost of calculating influence metrics for a node, even when running on a subset of the graph, we were not able to calculate the influence of all nodes. Thus, we compare values between our groups.

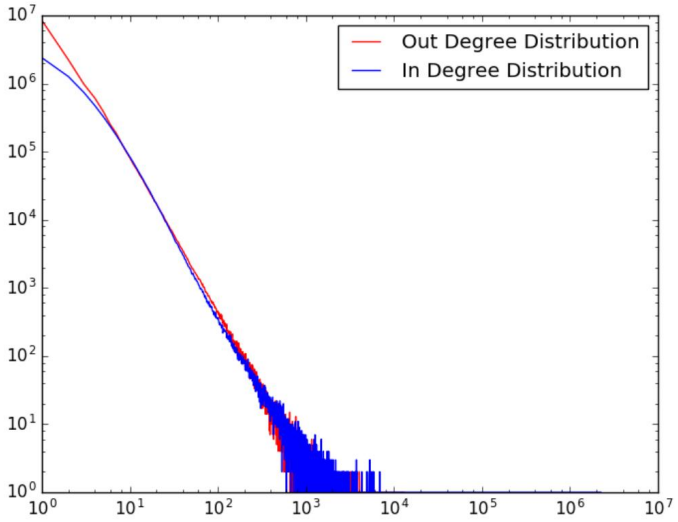

Figure 2: Degree Distribution of the Pinterest Dataset

5.1 Predicting Influence

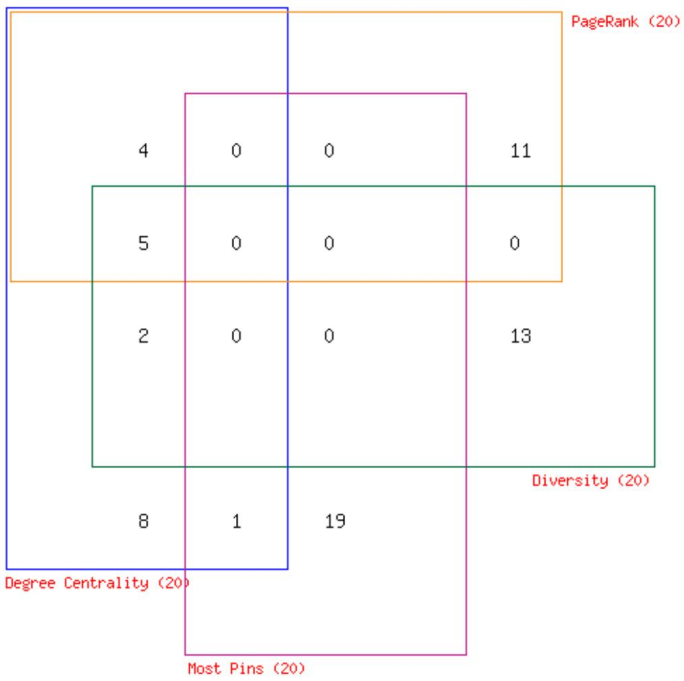

Figure 3: Overlap of top 20 nodes from each metric

Looking at the top 20 nodes produced by the top four metrics of PageRank, most pins, diversity, and degree centrality in Figure 3, we see a fair amount of overlap among them. Of interest, PageRank and diversity share 5 nodes that are also have the highest degree centrality. Most pins, on the other hand, while performing well, has almost no overlap with the rest of the metrics save for one node in the degree centrality top 20.

Furthermore, the top 100 nodes by number of pins has no overlap with either the PageRank or Diversity sets. This is not entirely surprising, as there is no reason prolific nodes would be 'diverse', but it does mean that the results found by these metrics are likely to be very independent. It also indicates that there may be a metric that takes content production into account while measuring

node importance metrics like PageRank.

5.2 Influence Metric Results

In order to evaluate the quality of the predictions made by the predictor metrics, we calculated the influence metrics on each node in each set. The metrics were calculated by tracking the node and its set of followers through the time-history of the network. We present top-1, top-5, and top-20 metrics for each set. The top-1 graph shows the influence metrics for the node with the single highest degree centrality, PageRank, etc. The top-5 graph shows the average influence for the nodes with the top 5 degree degree centrality, etc., and so on. One might consider a random sample of 20 nodes as our control group. Most nodes in the network have zero out-degree, so a random sampling of nodes results in zero influence metrics for all three categories.

Predictor Metric	Follow Uptake Probability	Follow Uptake Count	Re-Pin Count
Degree Centrality	0.02	0.51	3.87
PageRank	0.10	**3.41**	**5.65**
Most Boards	0.00	1.20	1.92
Most Pins	**0.14**	1.00	1.30
Diversity	0.04	2.11	5.23

Table 1: Influence Metrics for Top-1 by Predictor Metric

Predictor Metric	Follow Uptake Probability	Follow Uptake Count	Re-Pin Count
Degree Centrality	0.02	0.59	2.54
PageRank	0.03	1.38	**3.69**
Most Boards	0.02	2.84	2.89
Most Pins	0.04	**3.20**	3.09
Diversity	**0.11**	1.64	3.49

Table 2: Influence Metrics for Top-5 by Predictor Metric

Predictor Metric	Follow Uptake Probability	Follow Uptake Count	Re-Pin Count
Degree Centrality	0.05	1.03	2.39
PageRank	0.07	1.52	2.25
Most Boards	0.02	1.17	2.09
Most Pins	0.03	**1.98**	**3.17**
Diversity	**0.08**	1.73	2.45

Table 3: Influence Metrics for Top-20 by Predictor Metric

When only looking at the top predicted influencer by each predictor metric, PageRank identifies the best influencer, followed closely by diversity. It is difficult to differentiate top-1 performance from noise for these metrics, so we suggest looking at the top-5 or top-20 to get a sense of how each predictor metric performs.

As we begin to average the statistics across more of the top-k nodes identified by each predictor metric, Most Pins emerge as the most effective metric for identifying influencers with large impact on the network. This follows previous work in the field very closely, which has found that prolific nodes are very frequently highly influential in a network [8].

Interestingly, PageRank and diversity significantly outpace all of the other predictor metrics in identifying leaders with high follow uptake probabilities. Although these nodes have fewer followers on average than the prolific nodes, their followers copy their behaviours with much higher probablity. This signifies that PageRank and Diversity are successfully identifying properties of nodes that make them effective influencers. One possible explanation is that these nodes reach distinct areas of the network, so that a board, pin, or user unpopular in one community might be popular in another, and follow uptake probability captures whether at least one follower follows that object.

6 Conclusions and Future Work

We have demonstrated that there are many effective algorithms for quickly detecting influential nodes in a social network. Different algorithms often identify nodes with different properties and influence types, and we have shown which algorithms may be effective for different content types.

If an advertiser's main goal is to generate high numbers of clicks on some content, namely a pin, seeding the content to the most prolific nodes on the network may be the best way to propagate content. This result confirms previous work on smaller datasets, which have shown that prolific nodes are very frequently influential in a network. This also confirms findings that a large number of pins leads to more repins, and supports some evidence that more boards might negatively impact repins[5].

If an advertiser wishes to generate many follows out of some target neighborhood, selecting the nodes in that audience with high PageRank or Diversity scores would be the best choice among those tested. These algorithms, which try to reason about the structure of the network, are significantly more effective at identifying nodes that will effect a high percentage of their followers.

Additionally, if you imagine that advertisers are cost-constrained, it seems that given a very limited budget (i.e. only able to pay a single user), PageRank might be a good proxy. However, if a user is able to pay more users to seed content, number of pins becomes a more reliable metric.

It was also an interesting finding that the number of boards did not strongly predict any of our influence metrics, whereas number of pins did. This seems to confirms findings from Gilbert et al.'s research, where their linear regression model for calculating followers and repins had a negative coefficient for number of boards. One potential explanation for this might be that the effort required to have a large number of pins is much higher than the effort required to have a large number of boards, reflected in the quality and visibility of their content across the site.

There are several avenues for future work. First, other research can validate our findings in other networks or on different time slices of Pinterest. We wish to see whether PageRank or amount of content continues to be a strong predictor of node influence. Additionally, with larger computing resources, work should calculate the leadership metrics for a large portion of the graph, so comparisons can be made to the nodes with highest actual influence values, rather than comparisons between groups. Larger computing resources would also affect the runtime feasibility of Diversity over larger neighborhoods, potentially identifying better nodes for that metric. Second, future work would involve generating and testing new metrics, both in terms of predictor metrics or an influence metric that combines the three separate values used here, depending on the desired properties of the influence node. Third, rather than finding the single best predictor metric, research should attempt to combine other metrics into a model combining the values. For example, one promising direction might be able to bridge the gap between the effectiveness of the PageRank and Diversity metrics, and the raw impact of the most prolific nodes. There are is likely a middle ground between these two metrics that would result in more constant results across both uptake probability and uptake counts.

References

[1] Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining* (2011), ACM, pp. 65–74.

[2] Cario, J. *Pinterest Marketing: An Hour a Day.* Serious skills. Wiley, 2012.

[3] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. Measuring user influence in twitter: The million follower fallacy. *ICWSM 10*, 10-17 (2010), 30.

[4] Chang, S., Kumar, V., Gilbert, E., and Terveen, L. G. Specialization, homophily, and gender in a social curation site: Findings from pinterest. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work; Social Computing* (2014), CSCW '14, pp. 674–686.

[5] Gilbert, E., and Bakhshi, S. e. a. I need to try this!: A statistical overview of pinterest. *CHI 978*, 1 (2013), 2427–2436.

[6] Liu, L., Zhu, F., and Jiang, M. e. a. Mining diversity on social media networks. *Multimedia Tools and Applications 56*, 1 (2012), 179–205.

[7] Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[8] Theodoros Lappas, Evimaria Terzi, D. G. H. M. Finding effectors in social networks.

[9] Zubair Shafiq, Muhammad U. Ilyas, e. a. Identifying leaders and followers in online social networks. *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS/SUPPLEMENT 31*, 9 (2013), 618–628.