What Twitter Watches: Tracking Popularity Through Twitter Subgraphs

Francisco Cai, Blake Carpenter, David Philipson

December 8, 2010

1 Abstract

In our project, we investigated how popularity in society can be reflected in the online community. In particular, we examined how movies are represented on the Twitter social networks. We do this by analyzing a popular movies prevalence on Twitter network. For each movie, we construct a subgraph of the overall Twitter social graph based on discussion on Twitter, examine a number of properties of this graph, and relate it to the box-office performance of the movie. We obtain a surprising negative result and unexpected positive results for certain properties, and propose a hypothesis to explain them.

2 Introduction

In this paper, we look at how popularity in society can be reflected in the online community. In particular, we focus our attention on movies (popular and unpopular), and the discussion surrounding them on the online social network Twitter. We try to find some correlation between how much and how widely it is discussed on Twitter and the popularity of the movie in society. We hope to gain insight into whether we can predict the popularity or success of a movie based on what we can learn about it on Twitter. For this, we have collected tweets about movies on Twitter and for each movie, created sub-graphs of the Twitter social graph based on the relevant collected tweets. We then compare the sub-graphs for popular movies to the sub-graphs for the less popular movies. Before we discuss our methods and results, we will first explain the motivation behind choosing to use Twitter as a lens with which to assess the popularity, and why we chose to examine the popularity of movies specifically.

Present-day technology has enabled people spread their ideas and sentiments quickly and widely. Even in a world with email, blogs, Facebook, etc. Twitter stands out as a medium that is especially suited for impulsive and frequent broadcasts. In addition, tweets on Twitter are completely public, making data collection easier. For these reasons, Twitter is a great place to start analyzing the thoughts and sentiments of large group of people. We chose to examine popular movies because their popularity is easily quantifiable. For movies, we can use their box office

rankings as an intuitive measure for their popularity. In addition, movies are also topics that people like to discuss and share their experience with. If there is a popular movie with a unique, easily searchable name, we expect to find many users on Twitter writing about it. This paper describes the steps we took to analyze the structure of the Twitter discussion surrounding popular movies. We begin in the following section by exploring other relevant research done in this area and how it pertains to our project. In section 4 we discuss how we generated graphs that included users that tweeted about a movie, meaning they either have seen it, or at least have heard of it. Section 5 explains what properties we examined in these graphs and the methods that were done to better understand the structure of the Twitter discussion surrounding popular movies and movies. Section 6 presents that analyzed data and provides conclusions about our work.

3 Literature review

There has been many published papers analyzing the online social network Twitter. In "What is Twitter, a Social Network or a News Media?", authors Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon conduct a quantitative study.[1] One interesting result is that the Twitter universe differs from the real world social networks in certain ways. For one, the number of followers do no follow a power-law distribution, the degree of separation is smaller and there is less reciprocity (in terms of the 'following' relationship). This may implications for our project since we are trying to infer popularity in the real world from activity on Twitter.

In "Measuring User Influence in Twitter: The Million Follower Fallacy," the authors Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi empirically demonstrate the "Million Follower Fallacy", which states that a user who has many followers does not necessarily mean that he or she is influential.[2] While their paper addresses influence rather than popularity, their observation about the number of followers suggests that we would do well to look at a number of different aspects of the Twitter network when trying to infer popularity because the correlations may not be what we expect. For example, it may turn out that the number of tweets about a movie for example, may not give a good indication of its popularity or success in the box office.

In "Social Networks That Matter: Twitter Under the Microscope" authors B.A. Huberman, D.M. Romero, F. Wu discuss how the Twitter network is not representative of actual interactions between people, and the underneath this network of followers and followed lies a much sparser graph of social interactions.[3] Like the previous papers, this paper reveals another important difference between the structure and properties of the Twitter network versus real-life social network. We will need to keep this in mind when conducting our own research and interpreting the results of our research.

[1] Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. "What Is Twitter, a Social Network or News Media?" KAIST. Web. 09 Dec. 2010.

[2] Cha, Meeyoung, Hamed Haddidi, Fabricio Benevenuto, and Krishna P. Gummadi. "Measuring User Influence in Twitter: The Million Follower Fallacy." KAIST. Web. 8 Dec. 2010.

[3] Huberman, Bernardo A., Daniel M. Romero, and Fang Wu. "Social Networks That Matter: Twitter under the Microscope - Research Results - Social Computing Lab, HP Labs." HP Labs - Advanced Research at HP. Web. 09 Dec. 2010.

4 Data Collection and Construction

One main aspect of our project was extracting social graphs that represented all Twitter users who had either tweeted or received a tweet about a specific movie. This process was accomplished through the mining of multiple different datasets. One dataset contained 500 million tweets from June 2009 to December 2009.[4] Each tweet was contained information about date of tweet, content of tweet and username of the tweeter. The other data set contained the 1.5 billion relationships among Twitter users that were present during July 2009.[5] Because the first dataset dealt with usernames and the second dataset dealt with numeric IDs, another dataset was required that mapped user names to numeric IDs. Unfortunately, this mapping was not complete so not all usernames were able to be mapped to numeric IDs. As a result, many of the constructed graphs were missing links due to the inability to understand a users relationship in a graph without their numeric ID.

The movie data was collected using the two websites: Internet Movie Database (IMDB)[6] and BoxOfficeMojo.[7] We picked movies that were released during the time period of the Twitter dataset because movies reach maximum popularity soon after the time of release. We chose to use the movies release weekend box office rank as a metric for its true popularity. This was an obvious choice because the movies box office rank depends directly on its popularity. Each movie was paired with a few key phrases that would uniquely identify them in a tweet. This allows us to parse out tweets that directly discuss a movie.

Each movie corresponded to a relevant social graph that we constructed. These graphs had numeric ID's as nodes and edges between two nodes if one of the users had received a tweet discussing the movie from the other user. Our procedure for graph creation involved first extracting tweets that related to a movie. For each relevant tweet, we added an directed edge to a graph between the numeric ID's who made the tweet and every numeric ID's who received the tweet.

One of the main challenges of the project was parsing the Twitter data that was available to us. Extracting the relevant tweets required searching for movie keywords in more than 50 gigabytes of data, and this process had to be repeated for more than 50 movies. Once we had all the relevant tweets, to construct the graph structure, we had to map the usernames associated with the tweets to the corresponding numeric IDs. With the numeric IDs, we could then construct a graph structure using the user-follower relationships found in the Twitter social graph file. This part of the process required us to go through the roughly 1.5 billion lines in the social graph file for each movie, and save out the graph structure as a Networkx Graph object in Python. This required hundreds of computing hours, as well as gigabytes of free memory to store the Graph object before we save it to a Python pickle (a file containing the Python object in serialized form). Much of our time in the initial stages of this project was to set up this data processing and graph construction pipeline, and required us to find many workarounds. When analyzing very large data sets, the parsing of the data efficiently becomes a challenge in its own right.

In our analysis, we calculated many different properties of the graphs we had created and

[4]SNAP, http://snap.stanford.edu/data/twitter7.html

[5]KAIST, http://an.kaist.ac.kr/traces/WWW2010.html.

[6]The Internet Movie Database. Web. 09 Dec. 2010. ¡http://www.imdb.com/¿.

[7]Box Office Mojo. Web. 09 Dec. 2010. ¡http://boxofficemojo.com/¿.

formulated hypotheses about the relationship of these properties to the popularity of the movie associated with the graph. We wanted properties that had some intuitive real-world interpretation. Recall that the nodes are numeric ID's that identify a unique user, and edges are tweets between two users. First, we discuss the basic metrics we calculated:

- The number of nodes correspond directly to the number of users who have tweeted, or was following someone who tweeted about the movie.
- The number of edges correspond to the number of tweets related to the movie from one user to another.
- The average degree is exactly the average number of followers each user has.

We also looked at other properties, including the following:

- The number of connected components is an interesting property that can give insight into how much widespread appeal a movie has. All the graphs have many small connected components, and this reflects the fact that there are many groups of user and followers that in their own bubble, so to speak. This is made more likely by the fact that not all the usernames were mapped to numeric IDs when constructing the graph, so the graph may be more fragmented than it actually is. But for a graph to have more connected components than usual might suggest that it has limited appeal with certain groups rather than widespread appeal.
- In addition to the number of connected components, we also look at their average and maximum size. The average size might be skewed due to the large number of small connected component, but is easy to calculate and may yield unexpected information. The maximum size of a connected component may be directly related to how much a movie enjoys widespread appeal.
- The average core number of the nodes gives information about the connectivity of the graph, and hence, how connected the people tweeting about the movie are. This may be useful when trying to infer the popularity of a movie.

Given the sizes of these graphs (typically containing hundreds of millions of nodes), it was important for us to consider the time needed to calculate the properties we wanted to use, and factored into what properties we ended up being able to use. The above properties can be calculated in a reasonable amount time, since they require sub-quadratic time in the size of the graph. Other properties we wanted to look at, but did not because they would take too long to calculate, included the average clustering coefficient and the average shortest path length. Instead, we used faster heuristics as substitutes: Instead of calculating the average clustering coefficient, we looked at how the dense the graph was, e.g. the fraction of edges present out of total possible edges. Looking at this property could give us another way to estimate how much widespread appeal a movie has. One could expect a blockbuster that has widespread appeal to result in a Twitter graph with many far-flung nodes, and thus is less dense. To get a handle for the average shortest path, we selected

the node with the highest and calculated the average shortest path from that node. Thus, we only had to run the single-source shortest paths algorithm instead the all-pairs shortest paths algorithm. While this property probably has more variance than the normal average shortest path property, it may still yield useful information about how long chains of movie tweets are. It would not be surprising that longer chains of tweets are associated with more popular movies.

Even with our selection of properties that are relatively fast to calculate, doing so for these graphs took up to an hour each for many of the larger graphs. In sum, one run through the graphs where we calculate all the properties would take almost an entire day.

5 Results

Before computing the results, we predicted that the most significant predictor of success would be simply the volume of discussion on Twitter. After all, a movie which was widely discussed could be reasonably expected to also be widely watched, and likewise a movie which was seen by only a few viewers would likely have only a few people interested in discussing it as well. Hence, we predicted a strong positive correlation between the number of nodes in a movie's discussion graph and the success of that film.

We were thus very surprised to find that there is almost no correlation between the number of nodes and the success at all, regardless of whether we measure success by a movie's lifespan in theaters or by its overall domestic gross (fig. 5). Indeed, the plots visually appear similar to random noise, and the correlation coefficients of the number of nodes against lifespan and gross are respectively -0.02 and 0.012, so close to zero as to indicate no correlation between success and the number of nodes. Upon seeing the results of these plots, we decided to focus our analysis on

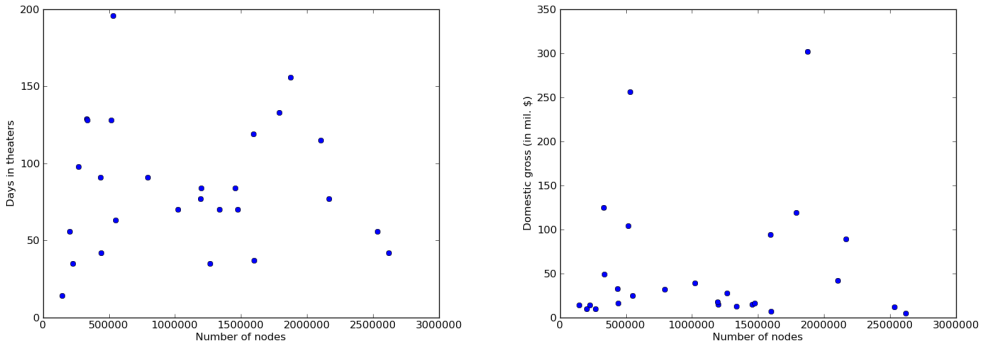

Figure 1: Movie success vs. size of discussion graph. The plots reveal no relationship. Left: success measured by lifespan. Right: success measured by domestic gross.

using a movie's time in theaters as an indicator of its success, rather than domestic gross. Both indicated the same information in the sense that movies with higher gross universally had longer lifespans as well, but using the lifespan gave plots which were easier to interpret. For example, it can be seen in figure 5 that many movies have very low domestic gross compared to a few hugely

successful movies whose gross eclipses the others, making the data harder to read than the more uniform lifespans.

The above result runs strongly counter to our intuition, as not even a weak positive correlation between the size of the discussion graphs and a movie's popularity was observed. Our surprise was compounded by the fact that most of the other properties we chose to examine showed similar irrelevance (fig. 2). As shown in the figure, there was no visible relationship between success and any of the six properties shown in the plot.

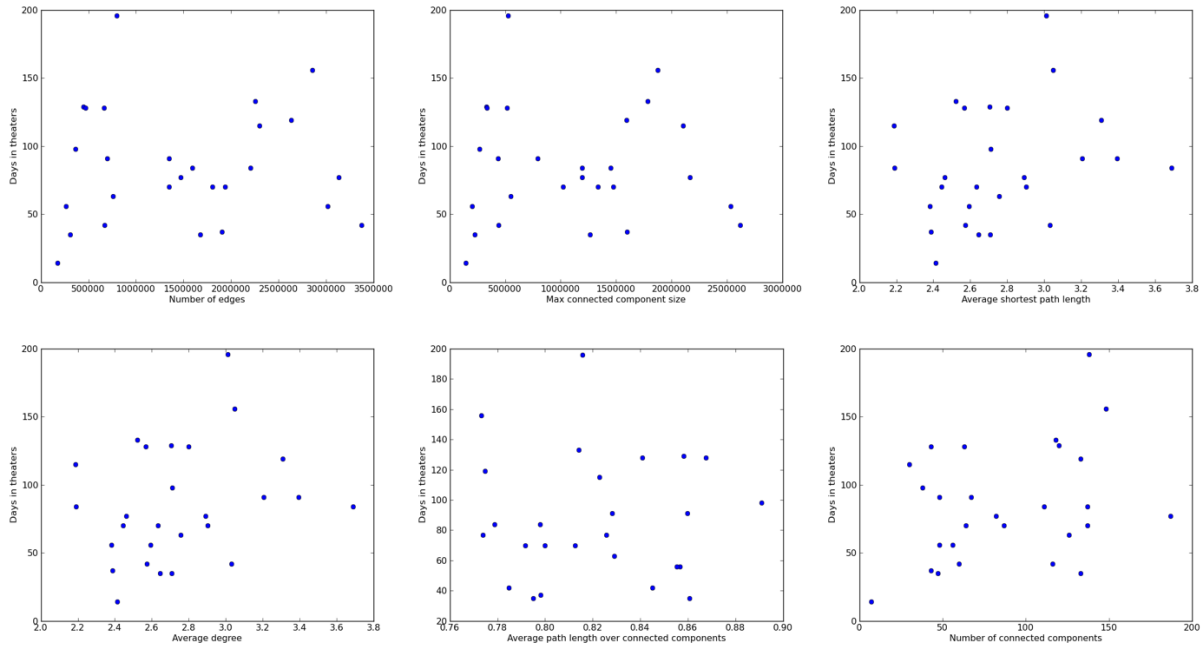

Figure 2: Lifespan vs. various properties. Top row: number of edges, maximum connected component size, average shortest path length. Bottom row: average degree, path length averaged over connected components, number of connected components.

However, we did find that two properties in particular demonstrated a notable negative correlation with a film's success, namely the density of the discussion graph and the average connected component size (fig. 3). These plots stood out from the others in that they are visibly less noisy. In particular, we can observe from the plots that there are no especially successful films at all whose discussion graphs have high edge density or many nodes per connected component on average. Our estimations from looking at the plots are backed up by computation. Looking at the correlation coefficients for each property against lifespan, we see that most of the coefficients are very close to zero, and that those for density and average connected component size are quite a bit larger in magnitude than the others (table 1).

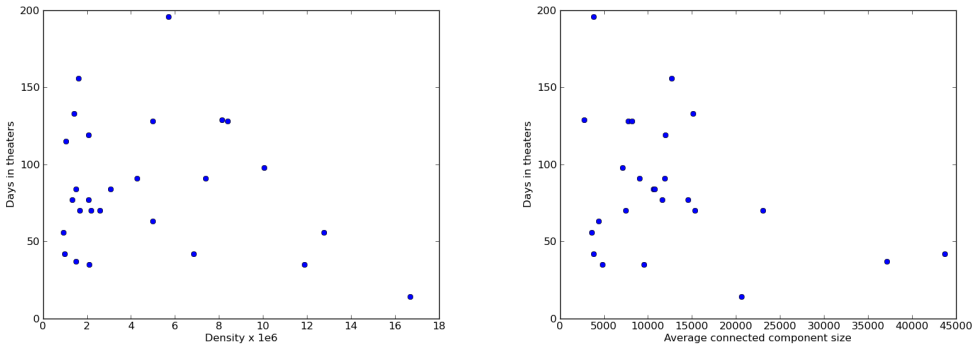

Figure 3: Properties with negative effects on success. Left: edge density. Right: average connected component size.

Number of nodes	-0.020
Density	**-0.47**
Average connected component size	**-0.36**
Number of edges	0.019
Average degree	0.17
Number of connected components	0.15
Max connected component size	0.019
Average path length	0.21
Path length average over connected components	0.004
Average core number	0.13

Table 1: Correlations between various properties against lifespan.

6 Analysis

We sought to find an explanation for the extremely unintuitive lack of correlation between a film and the volume of its discussion on Twitter, as well as the relative importance of the less obvious properties density and average connected component size. What could account for a widely popular and successful film receiving little discussion on Twitter? Conversely, how can a little-watched and unsuccessful film generate large amounts of Twitter discussion?

We put forward the following hypothesis: the size of a film's discussion graph–that is, the amount of discussion of the film on Twitter–is influenced most strongly by how much the audience of the film intersects with the core userbase of Twitter. After all, Twitter does not represent a uniformly random sampling of the moviegoing population in the style of a scientific survey. Rather, Twitter's userbase is a specific demographic, which may be much more interested in some topics than others. Then we might have a hugely successful movie which is not of interest to the typical Twitter user, or an unsuccessful movie which is targeted exactly to the sort of person who uses

Twitter, and we might see less discussion of the former than the latter.

Then what is the importance of edge density and average connected component size? Both of these are related to the idea of grouping. Higher edge density means that the graph is more filled in and we can expect more clusters of nodes which are more tightly connected than would be strictly necessary to connect them. A larger average connected component size indicates that users who are talking about a given film tend to fall into larger groups, rather than many separate islands of people. With the observation that an increase in either of these is related to a decrease in success, we interpret this as that a more spread out, distributed fanbase is beneficial to a film's success. We can think of these properties as looking through a window at a small part of a film's overall audience, and observing how tightly-knit or spread out that view is. This suggests that a clustering in the view that we get may be indicative of less wide-spread appeal, and thus harmful for a film's prospects.

As a side note, we realized that to investigate this hypothesis we would very much like to look at the clustering coefficients of the discussion graphs, which would seem to be another measure of how closely-connected a portion of a community is. Unfortunately, computing the clustering coefficients for all the graphs turned out to be computationally unfeasible; it would take $O(n^2)$ time, which is too much for these graphs with several million nodes and edges. The properties that we have will have to do.

	Bandslam	*Julie & Julia*	*Harry Potter*
Gross (in millions)	\$5	**\$94**	**\$302**
Lifespan (in days)	42	**119**	**156**
Number of nodes	**2,620,039**	1,592,706	**3,749,022**
Density $\times 10^{-6}$	**9.8**	2.1	1.6
Average connected component size	**43,667**	11,975	12,666

Table 2: Statistics for three case studies. Particularly large numbers are in bold.

To support our hypothesis, we looked at the numbers for three specific films as case studies (table 2). The first film was *Bandslam*, a particularly unsuccessful film which grossed a mere 5 million dollars and left theaters after a short 42 days. However, *Bandslam* generated a massive amount of discussion on Twitter. In fact, it had the largest discussion graph of any film during the investigated time period other than the huge blockbuster *Harry Potter and Half-Blood Prince*. To understand *Bandslam*'s success on Twitter, one need look no further than its target audience. *Bandslam* is a teen-musical starring Vanessa Hudgens, the star of the similar and hugely popular *High School Musical*. We can reasonably expect Twitter's active userbase to significantly intesect with the film's target audience of adolescents in middle- and high-school. Fitting with our hypothesis, *Bandslam*'s density and average connected component size were both high, as expected with its poor box office performance.

On the opposite end of the spectrum, we looked more closely at the film *Julie & Julia*, which performed well at the box office but received relatively little discussion on Twitter. *Julie & Julia* is a drama about Julia Childs starring Meryl Streep. Its target audience is generally older, which is most likely fairly disjoint from the main userbase of Twitter. The discussion graph density and average

connected component size are both small, which fits with our hypothesis given the film's success.

Finally, we looked at the massively successful *Harry Potter and the Half-Blood Prince*, which was popular both at the box office and on Twitter. As with *Bandslam*, we can expect a large intersection between the film's target audience of young adults and Twitter's userbase, which accounts for the large amount of discussion on Twitter. Once again, our hypothesis is reinforced by the fact that the discussion graph has small density and average connected component size.

7 Conclusions

The most important result of our research is that the real-world success of a film is not directly correlated to the volume of discussion of that film on Twitter. We find this result very surprising, and yet we can reach no other conclusion from the data.

We have also found that every property we could extract from the graphs, even the relevant ones (density and average connected component size) are too noisy to generate an accurate prediction for movie success from discussion graph topology alone. Nonetheless, there appears to be a clear relationship between the success of a movie and the properties discussed above, to the point where we found not a single example of a successful film with high density or average component size in its discussion graph.

We have arrived at the hypotheses described in the analysis section in our search to make sense of our results, but to confirm this hypothesis requires further study. Ideally, we would obtain demographic information about the Twitter userbase and the viewers of the assorted films, and we expect that we would find that the amount of intersection between the two is the most important factor in determining how much Twitter discussion appears. Such research would face the daunting task of collecting demographic data for a large number of films as well as of Twitter's userbase, but would be necessary to ultimately confirm our hypotheses.