# Empirical Study and Experiments on Information Virality Using Twitter Higgs Dataset

Zilong Wang (zilong@stanford.edu)
Zhiqing Zhang (zhiqing@stanford.edu)

## I. Introduction

Since the beginning of this century, the number of users consuming social medias has grown exponentially. According to Global Digital Report 2018, the number of social media users worldwide is 3.196 billion, representing more than 75% of the 4.021 billion internet users worldwide in 2018 [1]. Accompanying such tremendous growth is the gradual shift of social activities, marketing, advertising, and news consumption from offline to online as social networks provide the perfect medium of people, connected by similar backgrounds or interests, for information to spread. The intrinsic structure of social networks also influences the way information propagates. For example, news could spread differently on Facebook, where social connections are undirected, versus on Twitter, where connections are directed. The rise of social networks has made it easier to access information, but also brought along issues like fake news that sway public opinions much faster than traditional news media. To prevent fake news from virally spreading, we will first need to understand how news spreads across social network.

Our proposed project aims to exam how a scientific rumor [2] spreads across Twitter network. In this proposal, we summarize and critique three relevant papers on Twitter network analysis and Information Cascade. Our project will leverage and extend what is discussed in these papers to examine how information cascades spread across the network, identify communities and hub nodes, and explore the roles of social network, local community structure, and information cascade structure in viral outbreak..

## II. Literature Review

### A. *Social networks that matter: Twitter under the microscope*

Social networks on Twitter are often constructed from a list of declared followers and followees (i.e. people followed by a user) for analysis. In contrary with this popular practice, Huberman, Romero, and Wu investigate the underlying social networks constructed by the pattern of interactions that people have with their actual friends or acquaintances in this paper. Huberman et al. define a users friend as another user whom the user has directed at least two posts to (i.e. two @ interactions in posts or comments) and discover that on average, 90 percent of a users friends reciprocate attentions by being friends of the user as well. However, Twitter users have a very small number of friends compared to that of their followers and followees, implying a sparse network of actual friends underlying a dense network of followers/followees. They also show that this social attention is the key factor driving Twitter usage - the number of total posts saturates at less than 1000 as the number of followers increase while the number of total posts is positively correlated with the number of friends until it reaches a maximum point of 3201.

Huberman et al. provides some interesting insights on the effect of actual friend network underlying the follower-followee network on Twitter usage. We are planning on applying this hidden network analysis to our data set to see how friend network could contribute to the spread of news on Twitter. However, this paper solely emphasizes the social aspect of Twitter network and lacks analysis of Twitter usage as a source of information. In reality, Twitter is used by many people as a news source just as much as a social network.

There could exist users who have little interaction with friends yet still very active by proactively following public figures and sharing their posts and comments. We are interested in identifying these information hubs (public figures/accounts) and the follower communities around them to study how information flows into, across, and out of these communities.

Moreover, the definition of friend used in Huberman et al. appears to be another weakness. This paper defines friend as anyone who a user has directed a post to at least twice, which is too broad a definition in the context of real Twitter interactions. Any user could easily direct a post to a public figure (lets say New York Times) multiple times without being friends of each other. In our project, wed like to strengthen the definition of friend by requiring reciprocate following and explore how friend networks influence the spread of information across communities.

### *B. The Anatomy of a Scientific Rumor*

This article is the original publication on Nature that analyzes how the news of the discovery of a Higgs boson-like particle at CERN spreaded on Twitter and studies the spatial-temporal patterns of the information spreading across the network at local and global scale. The dataset and graphs of our proposal are also originally collected for this paper. The paper first provides Macroscopic and Microscopic plots of number of tweets in terms of inter-tweet time and inter-tweet space during different phases of the announcement. It unveils the bursty nature of user activities and refers to studies suggesting that spreading dynamics over complex network is influenced more by decision-based queuing processes and less sensitive to the overall network topology. The paper then further inspects the information spreading by modeling the dynamics of user activation first without user de-activation, then with deactivation. The author uses the assumption that neighborhood level correlation contribute little to the spread dynamics and uses a scale-free degree distribution to estimate if a non-active user is connected to active user. A large-scale data simulation is performed to validate this analytical model and a decaying activation rate is introduced for decreasing user interest over time in order to account for unexpected fast increase of active users in the beginning and rapid decrease after the announcement.

This article uses several brilliant modeling techniques to provide insights into spatio and mostly temporal patterns of the Higgs particle announcement, however we believe that there are many more aspects to be integrated in this network analysis that could potentially yield better understanding of the spread dynamics. First of all, the paper largely ignores the community structural topology and bases the analytical model on the assumption that local correlation plays a small part in spreading information across network. In reality, twitter networks tend to have many different subgraph structures. Some Twitter accounts tend to serve as a information hub node, connecting and broadcasting to a large number of nodes, while many other nodes tend to serve as information consumption nodes and rarely get retweeted. Secondly, while the paper provides a insightful view into how the news was spread in terms of number of tweets and active users, it would also be interesting to learn the topological pattern of the spread dynamics over time. Does the information tends to spread within tightly-related social communities faster, or does it reach broader audience first, then slowly saturate within the communities? Are there information hubs, or Centers of stars nodes and do they play an important role in spreading information across the network? We believe answering these questions with more diverse network analysis techniques can help us better understand the spread dynamics in this twitter network.

### *C. The Structural Virality of Online Diffusion*

To quantitatively understand how viral product or information diffuses, Goel et al. proposes a new measure of structural virality that interpolates between information that diffuses through one single large broadcasting and information that spreads through generations of relatively small-sized adaptations. By applying the concept of structural virality to the propagation of a variety of Twitter datasets, Goel et al. discovered that online popularity growth is made possible by a diverse combination of broadcasting and viral spreading,

but often times driven by the largest broadcast, which keeps the structural virality low.

We find the study of The Structural Virality of Online Diffusion a very interesting article. It provides many fundamental methodologies on how to model and study online diffusion, such as using structural diversity and structural virality as a continuous measurement of a diffusion tree. We would definitely try out structural virality analysis for our data as well, but we would love to take it further and compute a weighted structural virality index value by incorporating the retweet,reply and mention graphs together and assigning different weights to each edge according to corresponding feature set.

## III. Methods

### A. *Problem Statement*

The spreading dynamics of the discovery of Higgs particle through Twitter is a complex process. Does it follow a broadcast-type model, in which a popular node broadcasts to a large number of followers? Or does it more resemble a multi-layered viral spreading model, where the news travels through many Twitter friend circles to reach a large audience? In particular, how should we model and study the spread of information through large network systematically?

In this project, we want to answer the above questions by studying the Higgs information cascade diffusion trees in the context of the underlying twitter social network. Our initial empirical analysis will incorporate a wide range of features, such as the size, the root node degree, the average depth and the structural virality of the diffusion trees, and then closely examine how these features contribute to the overall size and structure of the information cascade. A microscopic examination of a few example cascades are visualized and analyzed using plotting tools. With the insight from our data analysis, we will further model how community structures influence the spread of information by labeling communities with Louvain algorithm. In our final experiment, we will augment the initial Retweet graph with additional weight and edges from all of the findings above, and perform cascade prediction on the new graph. The goal is that given a diffusion tree at certain time, our model will be able to predict whether the diffusion tree will keep the viral spreading pattern or not, given the features generated from the current tree.

The project contains the following components:

### B. *Dataset*

In this project, we took advantage of the Higgs Twitter Dataset [2] that is readily available on SNAP. This dataset was originally collected via the Twitter API by Domenico et al. before, during and after the announcement of the discovery of a Higgs boson-like particle at CERN between 1st and 7th July 2012. User activities that helped spreading this scientific rumor including retweeting, mentioning, and replying, were reported. From this data, four directional graphs of twitter activities and one directional graph of social relationships were extracted.

### C. *Network Properties*

The social network is a directed graph that reflects follower/followee relationships between active Twitter users who reacted to the discovery of Higgs particle. It is composed of 456626 nodes and 14855842 edges, with an average clustering coefficient of 0.1887.

The retweet network is a directed and weighted graph that represents retweet actions between users. It contains 256491 nodes and 328132 edges with an average clustering coefficient of 0.0156.

Like retweet, the reply network is another directed and weighted that reflects reply actions between users. There are 38918 nodes and 32523 edges in this network. Its average clustering coefficient is 0.0058.

Lastly, the mention network is a directed and weighted graph that represents the mention interactions between users. There exists 116408 nodes and 150818 edges in this network. Its average clustering coefficient is 0.0825.

### D. *Community Detection*

As mentioned in the problem statement above, in order to understand how social communities play their roles in the spreading of information, we

would like to first explore the kind of communities that exist in the twitter social network. Due to the nature of social relationships, we expect that community structures exist intrinsically in the social network. For this classic community detection problem, we adopted the Louvain algorithm as it provides a fast means to detect communities in a large network. Louvain algorithm is a greedy optimization method that aims to optimize the modularity of a partition of the network.

The modularity of a partition is defined as:

$$Q = \frac{1}{2M} \sum_{i,j} \left[ A_{ij} - \frac{k_i k_j}{2M} \right] \delta(c_i, c_j) \qquad (1)$$

where $A_{i,j}$ represents the weight of the edge between $i$ and $j$, $k_i = \sum_j A_{i,j}$ represents the sum of the weights of the edges attached to vertex $i$, $c_i$ is the community to which vertex $i$ is assigned, the $\delta$ function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise and $2M = \sum_{i,j} A_{ij}$[5].

As described in Blondel et al., each pass of the Louvain algorithm contains two phases:

*1) Phase 1*

- Start with each node in its own community.
- For each node i, loop through all its neighbors j and evaluate the gain in modularity by removing i from its current community and placing it in js community. Node i is then placed in the community that maximizes gain in modularity.
- If no positive gain is possible, i remains in the original community.
- Repeat the above process for each node until no further improvement is possible.

*2) Phase 2*

- Contract the original graph G to a new graph H by making each community found in Phase 1 and node. The weights of edges between two nodes in H is equivalent to the sum of the weight of edges between the corresponding two communities in G. The sum of weights of edges within a community in G is converted to a self-edge of the same weight in H.

Repeat Phase 1 and 2 until no improvement in modularity is possible, which means optimization is reached.

For this project, instead of implementing the Louvain algorithm from scratch, we had used the community detection module for NetworkX to achieve this goal.

### *E. Diffusion Tree*

In this project, diffusion tree is used to model the cascade of information through twitter. We first generated a list of diffusion trees using the twitter Higgs time-activity data. Each row in the dataset represents a diffusion event (edge) and contains the "from" (source node) and "to" (target node) user IDs, the interaction type (retweet, mention, or reply), and the timestamp. If the source node of any event has never been referenced in any previous diffusion event, we then define it as a "seed node", which serves as the root node of a new diffusion tree. We also made the assumption that the diffusion tree that a node belongs to does not change once it is first set, though it can be the target of diffusion events from other diffusion trees. By iterating through the Higgs time-activity data, diffusion trees are generated and grown in a temporal order. By the end of this processing, four sets of diffusion trees are generated for retweet, mention, reply, and all activities. Note that diffusion trees with less than 2 nodes are filtered out to reduce noise.

### *F. Structural Virality*

There exists various ways to define the virality of a graph. For this project, we adopted the definition of structural virality $v(T)$, discussed in the Goel et al. paper, as the average shortest distance between all pairs of nodes in a diffusion tree $T$:

$$v(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} \qquad (2)$$

where $dij$ denotes the length of the shortest path between nodes $i$ and $j$. This definition also known as Wiener index and provides a continuous measure of structural virality. The higher the value of $v(T)$ is, the farther apart the adopters are from each other in the cascade, thus suggesting an viral diffusion event deep into many layers of nodes. On the other hand, a lower value of $v(T)$ generally

**Algorithm 1** (Computing $v(T)$)
**Require:** $T$ is a tree rooted at node $r$

**function** SUBTREE-MOMENTS($T, r$)
  **if** $T$.size() = 1 **then** ▷ The base case
    size ← 1
    sum-sizes ← 1
    sum-sizes-sqr ← 1
  **else** ▷ Recurse over the children of the root $r$
    **for** $c \in r$.children() **do**
      $\text{size}_c$, $\text{sum-sizes}_c$, $\text{sum-sizes-sqr}_c$ ← SUBTREE-MOMENTS($T, c$)
    size ← 0
    sum-sizes ← 0
    sum-sizes-sqr ← 0
    **for** $c \in r$.children() **do**
      size ← size + $\text{size}_c$
      sum-sizes ← sum-sizes + $\text{sum-sizes}_c$
      sum-sizes-sqr ← sum-sizes-sqr + $\text{sum-sizes-sqr}_c$
    size ← size + 1
    sum-sizes ← sum-sizes + size
    sum-sizes-sqr ← sum-sizes-sqr + $\text{size}^2$
  **return** size, sum-sizes, sum-sizes-sqr
**function** AVERAGE-DISTANCE($T, r$)
  size, sum-sizes, sum-sizes-sqr ← SUBTREE-MOMENTS($T, r$)
  $\text{dist}_{\text{avg}}$ ← $[2 \cdot \text{size}/(\text{size}-1)] \times [\text{sum-sizes}/\text{size} - \text{sum-sizes-sqr}/\text{size}^2]$
  **return** $\text{dist}_{\text{avg}}$

Fig. 1. Algorithm for computing $v(T)$ by Goel et al.

suggests a less viral event, thus more resembles a noisy broadcast model.

The algorithm for calculating $v(T)$ proposed by Goel et al is represented by Fig. 1. In our project, we took advantage of an API for calculating average length of shortest path of a graph conveniently provided by NetworkX.

### G. *Outbreak Detection*

In this section, we further explored the idea of predicting if a certain diffusion tree will go viral given its current characteristics. We generated 2 sets of features, one derived from the diffusion tree and its corresponding social graph, another from the community detection results. The aim is to compare whether adding additional feature sets based on community labels can further improve our models prediction accuracy, thus arguing how important local communities play an role in the diffusing process.

To generate the labels, we selected a time

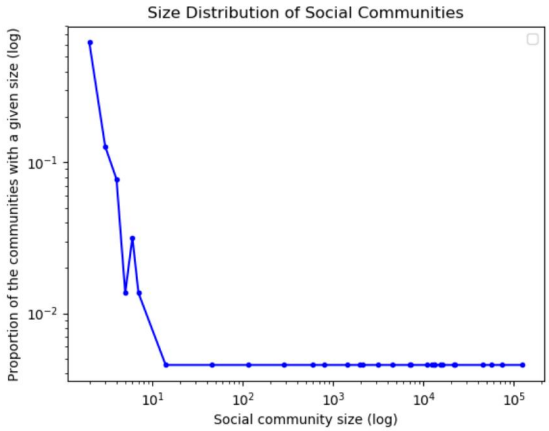


Fig. 2. Size distribution for communities detected from Social network

stamp (1341403280) which occurred shortly after the main news broke out on Twitter and take a snapshot of all the diffusion trees at this point of time. We then compare each diffusion tree with its final form at the end of the data collection. If the diffusion tree has doubled its size in the given time window, we will consider it to be viral and give it a label of 1. Otherwise, the tree is given a label of 0.

Once we generated the necessary data feature sets and corresponding labels, we randomly split the data into 90% for training and for 10% testing, and then we feed the data into a Random Forest classifier with 100 trees to train our model.

*1) Feature Set*

The first 5 columns of feature set without community labels (Table II) and of with community labels (Table III) can be viewed below.

*a) ID*

ID of the Diffusion Tree. Each diffusion tree will have a unique id. The tree id is only for identification purpose and it is not part of the feature set.

*b) Root Deg*

The degree of the root node of a diffusion tree. A high root node degree in diffusion tree usually indicates a broadcast diffusion model.

*c) Root Deg Social G*

The degree of diffusion tree's root node in the original social graph.

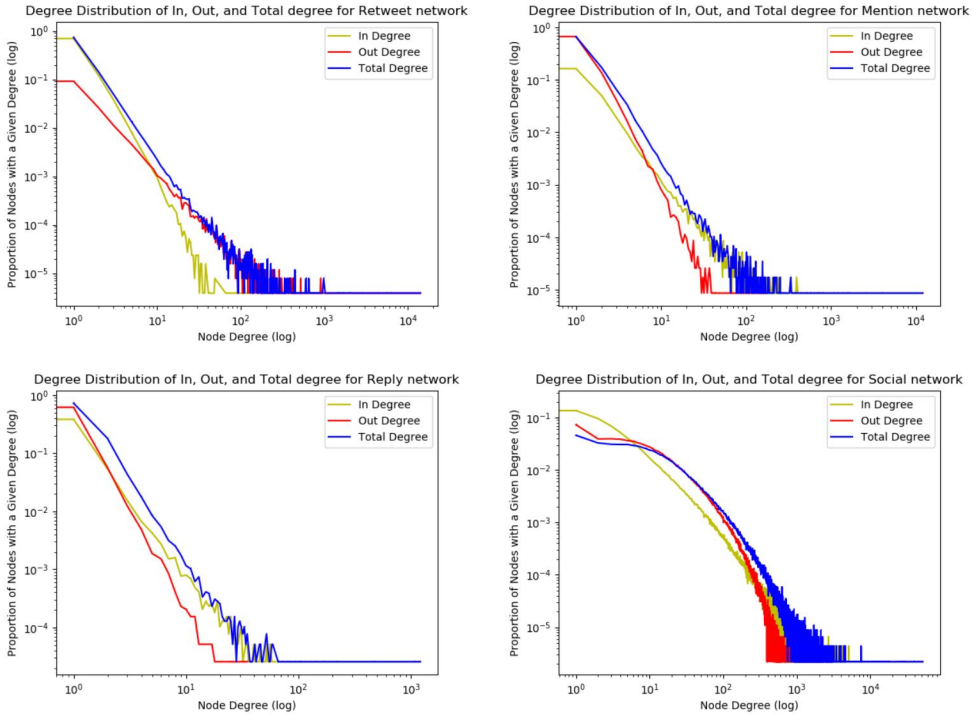


Fig. 3. Degree distributions for Retweet, Mention, Reply, and Social networks

*d) Structural Virality*

The structural virality value of the diffusion tree

*e) Node Cnt*

The total number of nodes in the diffusion tree.

*f) Edge Cnt*

The total number of edges in the diffusion tree.

*g) Root Community Size*

The size of the community that the root node belongs to.

*h) Largest Community Density*

Representation of the density of the largest community within the current diffusion tree. This feature measures if the diffusion tree is currently dominated by a certain community or evenly spread through several different groups. The value of the largest community density is calculated by:

$$X_i = \frac{C_i^2}{N_i} \tag{3}$$

where $C_i$ represents the largest count of appearance of a single community within the diffusion tree, and $N_i$ is the total number of nodes of the community.

## IV. Results and Discussions

### A. *Summary Degree Distribution*

Our first step is to generate the degree distribution of the four networks to get an idea of what we can expect from the data and how the networks are distributed. As shown by the graphs above, the in, out, and total node degree distributions for social, retweet, reply, and mention networks all follow a similar power-law distribution, which is expected in real-world networks.

The social network has some of the largest node degrees, which makes sense intuitively as the degree for each node in the retweet, reply and mention graph cannot exceed that of the social graph. Among the three twitter interactions graphs, retweet and mention graphs have the larger distribution of nodes with high degree, whereas the reply graph hosts more nodes with small degree.

### B. *Community Size Distribution*

Since the Louvain community detection algorithm assumes undirected graph, we had to convert these twitter interaction graphs

to undirected before performing community detection. 220 communities were successfully identified from the twitter social network. Fig. 2 represents the distribution of communities detected by the Louvain algorithm. Of these 220 communities, the largest community contains 125344 nodes, whereas the smallest community contains only 2 nodes. The mean community size is 2075, the mode and median community size are both 2 - there are 137 communities with a size of 2 and large communities are very few.

### C. *Traits of the Diffusion Trees*

Our next step is to generate the list of diffusion trees for each graph, and then plot the fundamental traits such as degree distribution, average depth and cascade size for each graph. Fig. 4 shows two example diffusion trees (19271 and 928) from the all activity network.

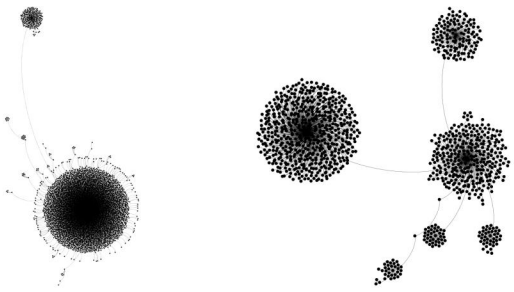

Fig. 4. Tree 19271 and 928 constructed from total activity data set

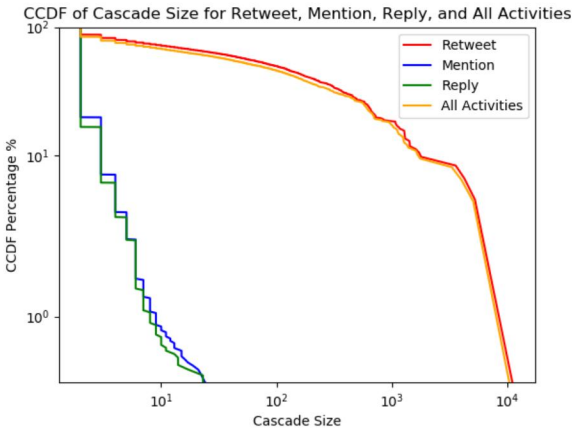


Fig. 5. CCDF of cascade size for Retweet, Mention, Reply, and All Activities

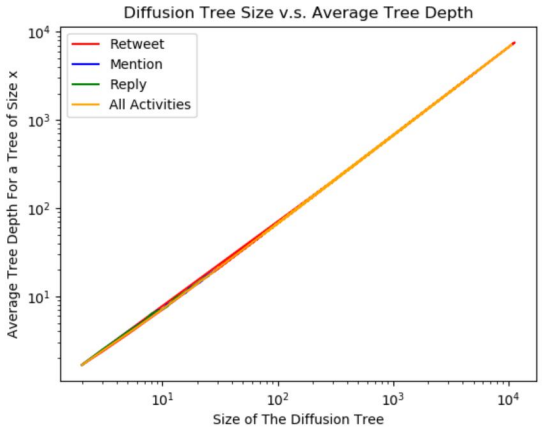


Fig. 6. CCDF of cascade size for Retweet, Mention, Reply, and All Activities

#### 1) *Cascade Size*

The loglog plot of Fig. 5 gives us an overview of the sizes of cascade trees for each graph. The All Activity and Retweet networks have the largest cascade in terms of size, where as Mention and Reply networks have relatively small cascades.

This makes sense intuitively as retweeting can be regarded as the strongest among the three in terms of reaching out to a broad audience, as it is the main catalyst of information cascade. It is not surprising that the All Activity curve closely resembles the Retweet curve.

#### 2) *Average Depth*

Figure 6 shows high correlation between the average depth of diffusion trees and the size of the diffusion trees for all four networks. This tells us that larger diffusion tree often means information diffuses to deeper layers, possibly inferring that as the information cascade grows bigger, it demonstrates more characteristics of viral diffusion than a broadcast. Otherwise, the average depth will be flatter, indicating that the diffusion tree size is more correlated to the degree of celebrity nodes.

#### 3) *Root Node Degree*

There are studies that focused on utilizing features of the root node as key indicator, and the conclusion is that celebrity nodes with a higher degree reaches larger audience and thus spreads further. Users with large follower counts on Twitter

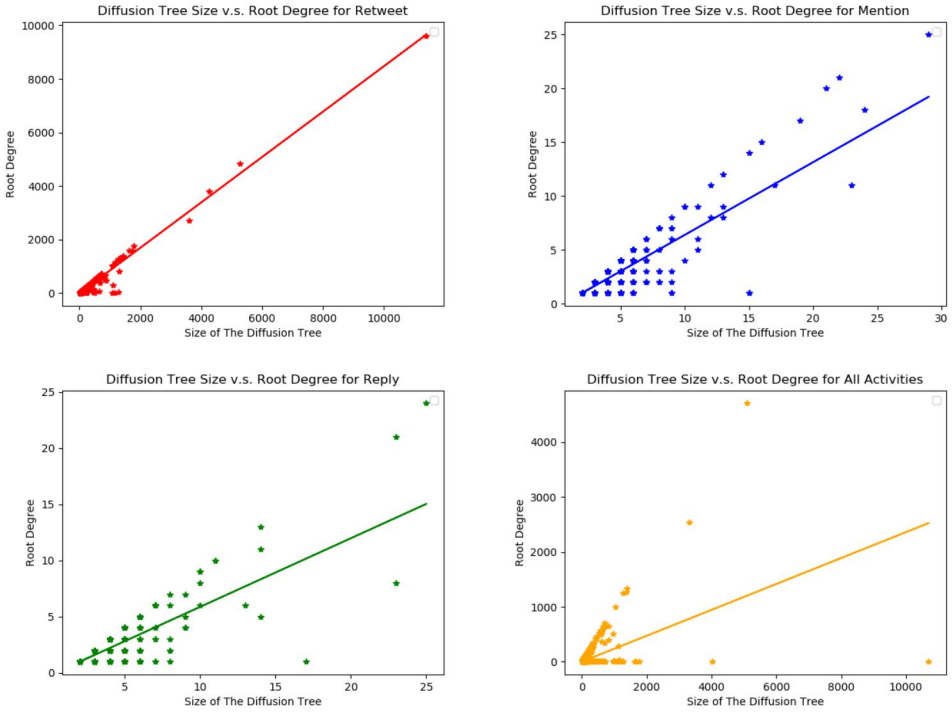


Fig. 7. Cascade size v.s. Total root degree in Retweet, Mention, Reply, and All Activities

generated the largest retweet cascades.[6] By plotting the relationship between tree size and the root node degree (Fig. 7), we hope that it can confirm our hypothesis from 2) above that the spread of the Higgs particle news resembles more of a viral diffusion.

After examining the plots above, we have drawn some interesting insights:

- The cascading dynamics is a mix of viral diffusion and broadcast diffusion. While all four graphs exhibit similar trends, the All Activity plot appears to support the mixed cascading dynamics most significantly. In the All Activity plot, we see a clear separation of two types of trees - those with high root degree but relatively low cascade size compared to the best fitting line, indicating a broadcast diffusion, and those with low root degree but high cascade size, a classic viral diffusion. Compared to the All Activity plot, the Retweet plot is also showing a mixture of the two dynamics, but less significantly. There is no cascade that grows virally to over 10000 nodes by retweet only, as we do see in all activities. Mention and Reply plots are relatively sparse, which is expected as they have fewer data points, but they do show a mixture of both as well.
- By plotting the linear regression of of each graphs data, we see a general upward correlation between the root node degree and cascade size. This observation lies with the previous finding that the spreading dynamics for Higgs news is a mixture of broadcast and viral diffusion events.

### D. *Structural Virality*

In Fig. 8, we plotted the complementary cumulative distribution function of structural virality for each graph. This graph reinforces our previous finding that structural virality varies from about 1 to 5, indicating that the diffusion trees are a mix of short/wide broadcast trees as well some as deep/spread-out word-of-mouth trees. We also have the similar observation as in study [5], that there is no bi-modal distribution structure for structural virality. This indicates that there is no tipping point for diffusion, compared to some of the classic theories. Another interesting observation here is

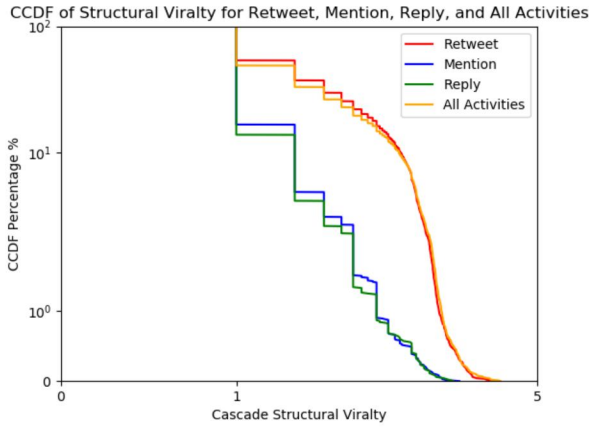


Fig. 8. CCDF of Structural Virality for Retweet, Mention, Reply, and All Activities

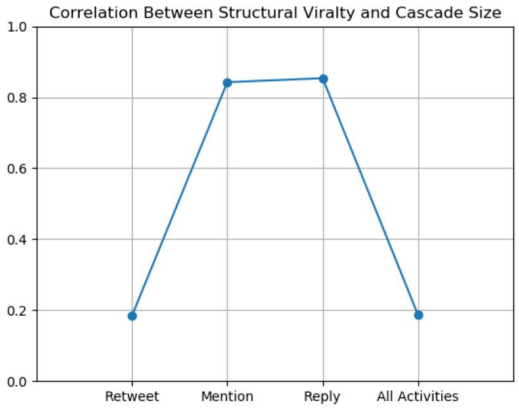


Fig. 9. Correlation between Structural Virality and Cascade Size

that Twitter mention is slightly more viral than retweet, which could suggest that the inner social network and close communities play a crucial role in information cascading.

Figure 9 illustrates the correlation coefficient between structural virality and the size of cascade. As we can see, all four graphs show a positive correlation between size and virality, indicating that the major driving force behind cascade growth is likely word-of-mouth diffusion. Among the four graphs, Reply graph has the highest correlation coefficient, closely followed by the Mention graph. All activities and Retweet graph shared the same low correlation, indicating that high cascade size does not necessarily infer high structural virality, and the total activity is dominated by retweet.

### E. *Outbreak Prediction*

Table I represents the results of the prediction while Table II and III show the values of the feature set. As seen from Table I, we do not have a highly accurate classifier to detect outbreak. the f scores are low in both models, likely because of the lack of sufficient data and effective features for training. A slight improvement in accuracy is obtained by adding the community features into the training set, however it is too small to have any statistical significance. This result corresponds to our previous observation that most viral diffusion events are dominated by broadcasting model rather than from one local community to another local community.

From this experiment, we can conclude that viral spreading from one local community to another local community does not contribute much to increasing the chance that a diffusion event to become even more viral, although it does happen infrequently. Therefore, including the community detection derived features does not significantly improve our prediction accuracy.

## V. Limitation and Future Work

*1) Limitation*

*a)*

First limitation we run into is that the majority of the diffusion trees only end up having 1 or 2 edges. Therefore, the majority of the diffusion trees will be classified as ”non-viral”, resulting a high bias in our classification model. To combat this problem, we experimented with the idea of a minimal tree size threshold, which filters out diffusion trees that contain fewer than a certain number of nodes (5) in our case. While this approach is effective in reducing noise in the data, it greatly reduced our sample size from 40k 5k, leaving us susceptible to high variance and over-fitting. We used Random Forest algorithm with 100 500 trees to compensate for this problem, but the improvement is not significant.

TABLE I

PREDICTION ACCURACY FOR TEST DATA

| Feature Set | Accuracy | Weighted Avg f1-score |
|---|---|---|
| Without Community | 0.6218487394957983 | 0.61 |
| With Community | 0.6659663865546218 | 0.65 |

TABLE II

FEATURE SET EXAMPLE WITHOUT COMMUNITY FEATURES

| ID | Root Deg | Root Deg Social G | Structural Virality | Node Cnt | Edge Cnt |
|---|---|---|---|---|---|
| 0 | 1.0 | 65.0 | 1.000000 | 2.0 | 1.0 |
| 1 | 14.0 | 659.0 | 1.866667 | 15.0 | 14.0 |
| 2 | 1.0 | 147.0 | 1.000000 | 2.0 | 1.0 |
| 3 | 72.0 | 2283.0 | 2.098568 | 78.0 | 77.0 |
| 4 | 605.0 | 32160.0 | 2.309795 | 717.0 | 716.0 |

TABLE III

FEATURE SET EXAMPLE WITH COMMUNITY FEATURES

| ID | Root Deg | Root Deg Social G | Structural Virality | Node Count | Edge Cnt | Root Community Size | Largest Community Density |
|---|---|---|---|---|---|---|---|
| 0 | 1.0 | 65.0 | 1.000000 | 2.0 | 1.0 | 45912.0 | 60.0 |
| 1 | 14.0 | 659.0 | 1.866667 | 15.0 | 14.0 | 125344.0 | 9.0 |
| 2 | 1.0 | 147.0 | 1.000000 | 2.0 | 1.0 | 125344.0 | 0.0 |
| 3 | 72.0 | 2283.0 | 2.098568 | 78.0 | 77.0 | 125344.0 | 3.0 |
| 4 | 605.0 | 32160.0 | 2.309795 | 717.0 | 716.0 | 125344.0 | 0.0 |

*b)*

Another limitation in our prediction machine learning algorithm is choosing the specific time stamp for the ”snapshot” and then defining the threshold for virality. Choosing a time stamp too early would mean most of the diffusion trees still have only 2 to 3 nodes and as a result, most of them will double its size by the end of the data collection period. In this scenario, we will need a high threshold for virality. For example, only diffusion trees that quadruple its size in the end can be labeled as viral. On the other hand, choosing a time stamp too late would mean that most of the diffusion trees have already passed the viral spreading period, thus they are unlikely to continue doubling the tree size. In this scenario, a smaller virality threshold is required to reduce bias in the model. We have performed dozens of experiments with different combinations of threshold, and we are unable to consistently perform over an accuracy of 0.75. This result implies that outbreak is a complex and random process, and it is indeed very difficult to have a generic and accurate model to predict whether an event will go viral.

*c)*

We also have a limited feature set. The intention here is that by explicitly selecting features, we can mitigate some of the impact of small data problem. However, effective feature selection requires deep domain expertise in the field, and it is prone to humane subjectivity. Even though we tried our best to select features we think are essential in outbreak detection, it is far from perfect and can be improved with further study of the subject.

## APPENDIX

Github Repo: https://github.com/zhiqing1993/cs224w-project

## REFERENCES


[1] https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research

[2] M. De Domenico, A. Lima, P. Mougel and M. Musolesi. The Anatomy of a Scientific Rumor. (Nature Open Access) Scientific Reports 3, 2980 (2013).

[3] B.A. Huberman, D.M. Romero, F. Wu. Social networks that matter: Twitter under the microscope. First Monday, 14(1), 2009.

[4] Goel, S., Anderson, A., Hofman, J.M., Watts, D.J. (2016). The Structural Virality of Online Diffusion. Management Science, 62, 180-196.

[5] Blondel, Vincent Guillaume, Jean-Loup Lambiotte, Renaud Lefebvre, Etienne. (2008). Fast Unfolding of Communities in Large Networks. Journal of Statistical Mechanics Theory and Experiment. 2008.

[6] Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyones an influencer: Quantifying influence on twitter. Proc. Fourth ACM Internat. Conf. Web Search and Data Mining (Association for Computing Machinery, New York), 6574.