# A Study of Meme Propagation: Statistics, Rates, Authorities, and Spread

Onkar Dalal
Stanford University
496 Lomita Mall
Stanford, CA, 94305
onkar@stanford.edu

Deepa Mahajan
Stanford University
450 Serra Mall
Stanford, CA, 94305
dmahajan@stanford.edu

Ilana Segall
Stanford University
496 Lomita Mall
Stanford, CA, 94305
isegall@stanford.edu

Meghana Vishvanath
Stanford University
496 Lomita Mall
Stanford, CA, 94305
mvishvanath@stanford.edu

## ABSTRACT

We use the Meme-tracking system to understand how phrases spread across news sources [3]. We have observed how a meme gains momentum and loses popularity, and observed how the category of the meme affects its life. We begin by looking at various statistics from the frequency data to understand differences between the lifetime of memes across subject matter and media source, and look for methods to predict how the shape of the graph will change over time. Additionally, we consider the graph of news sources and directed edges from the sources that are hyperlinked in the original article. Here we find the top authoritative news sources and blogs. Additionally, we built a theoretical example to determine the actual influence network. Finally, we consider a cascade network on this graph to determine whether the SIS model is reasonable to model meme propogration.

## Keywords

Meme-tracking, Blogs, News media, News cycle, Information Diffusion, Social Networks

## 1. INTRODUCTION

Social networks have become increasingly popular and studied over recent years. In addition to the typical friendship networks (such as Facebook or MySpace), social networks have emerged in various fields: photo sharing, online game playing, and even the news media. We seek to explore the process of news information diffusion through the study of meme propagation. In this paper, a meme represents a phrase that travels through the news blog network [4]. We hope to observe key characteristics of the propagation in order to understand and appropriately model the propagation of memes in the online news cycle.

## 2. BODY

We begin with a description of the data we used and techniques we implemented in order to understand our data. The Method section is broken into two subsections: "Frequency Data" and "Time Data". The first represents data of meme occurrence over time and has no information regarding which sources used the meme. The second represents a multigraph where nodes are the new sources and edges correspond to whether those nodes share the use of a meme. Section 3 refers to our theoretical construct and describes how we would visualize the real influence network.

### 2.1 Data

We look at the MemeTracker data which tracks memes over various news media and blog sources during August 2008 until April 2009 [3] . We used both the phrase cluster data and raw phrases data for the month of November 2008, provided on the website.

### 2.2 Method

#### 2.2.1 *Frequency Data*

Our first goal was to understand how the magnitude of posts containing a certain meme was changing over time, and see if there were any statistics that might help us understand the nature of the media cycle. We were interested in seeing if there were differences in the key features of the meme propagation of various types of media (blogs and news sources) and different categories of news (politics, entertainment, etc).

Initially, we implemented a Python module to produce the frequency of a phrase over any user-specified time step (following the frequency over time model that was used in [4] ). In order to characterize the frequency patterns we saw in the meme data we looked at five key statistics: peak value, percentage of lifetime elapsed before peak is reached, lifetime, diameter of the peak (distance from start to end of the peak hump), and ratio between the rate of ascent and descent of the peak. For the purposes of this analysis, we considered only the highest peak for each graph, and considered the hump to begin and end when the number of posts in a time period was greater than and less than 10% of the magnitude of the peak, respectively. We used a rough approximation for the relative ascent and descent rates by calculating the absolute value of the ratio of the rate of ascent divided by the rate of descent. Rate of ascent was calculated by

$$\frac{(\text{peak} - \text{starting ascent point})_{\text{FREQ}}}{(\text{peak} - \text{starting ascent point})_{\text{TIME}}}$$

Rate of descent was similarly calculated.

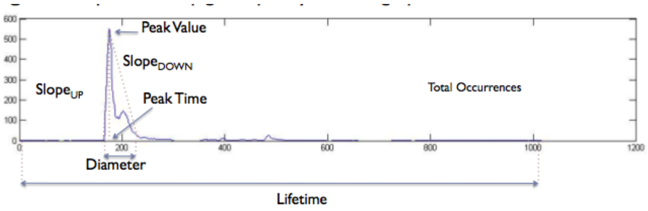


**Figure 1: Gathering features from the scatterplot of frequency data for the phrase "lipstick on a pig"**

We looked at the key statistics for each meme and then looked at categories of memes and the average values for these five statistics pointed out in Figure 1. We then compiled the statistics on each of these data sets independently.

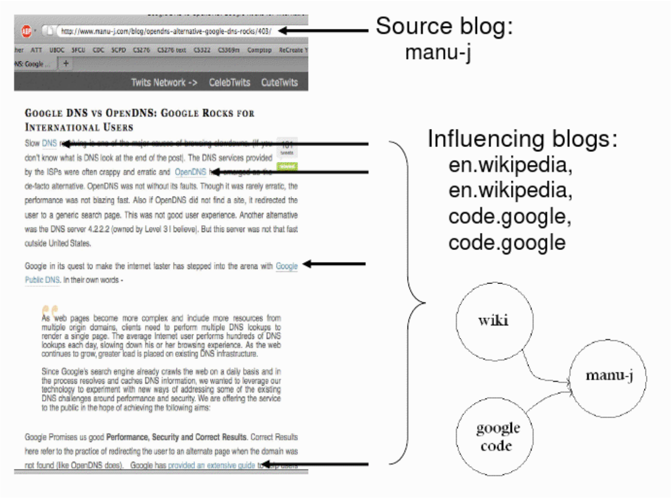
We utilized a t-test with a 95% significance threshold. This process was run on several different data sets that were created by organizing and dividing up the original cumulative data into conceptual blocks. In our first comparison, we took each meme and ran the frequency module on the data to separate the occurrences due to blogs (Blog Data) and the occurrences due to the mainstream media (Mainstream Data). We also hand categorized the first 200 memes into five categories, (**E**)ntertainment, (**P**)olitics, (**N**)onpolitical, (**O**)ther, and (**T**)echnology, and treated each category as a new data set. Lastly, we subdivided the Politics category into quotes made by the Republican political party and quotes made by the Democratic political party.

### 2.2.2 *Time Data*

We generated the following multigraph on a small subset of phrase cluster data. Each node refers to a news media or blog source (and is specifically labeled as a tuple of (name-of-source, **B** or **M**) referring to **B** for blog and **M** for news media source). If two sources use the same root phrase, they share an edge labeled with that phrase. When the graph is constructed, the number of edges between two nodes is the number of their shared root phrases. Although this is represented as a graph, the simplified structure of this is a set of levels, where the $n^{\text{th}}$ -level connects entirely with every node in levels 1 to $n$ (this basically forms a complete graph on every node that uses that phrase). On a small subset of the phrases provided in MemeTracker data.

Our next goal was to understand the influence network of the media-blog network. For this we used the raw phrase cluster data of the month of November 2008. A node in this graph is a news source or blog source, and there is a link from source **a** to source **b** if source **b** cited a link in source a. See Figure 2 for more details. This generated a graph with 692,209 nodes and 2,189,909 edges.

The rationale behind this construction is that this gives us the best approximation of the inference network on the induced hyperlink structure. We make the assumption that if a blog cites a bunch of sources it can reasonably model the behavior that the author read those sources and decided to write his own article on it, with the influence of the sources he read. In this way, we draw a graph of influencing sources across our network. Naturally, this technique is not always correct because news sources will frequently link to many other articles inside their own domain in the form of "Related Articles." We account for this over-counting by allowing one source to only have unique influencing sources.


http://www.manu-j.com/blog/opendns-alternative-google-dns-rocks/403/

GOOGLE DNS VS OPENDNS: GOOGLE ROCKS FOR INTERNATIONAL USERS

Slow DNS resolving is one of the major causes of browsing slowdowns. (If you don't know what is DNS look at the end of the post). The DNS services provided by the ISPs were often crappy and erratic and OpenDNS has emerged as the de-facto alternative. OpenDNS was not without its faults. Though it was rarely erratic, the performance was not blazing fast. Also if OpenDNS did not find a site, it redirected the user to a generic search page. This was not good user experience. Another alternative was the DNS server 4.2.2.2 (owned by Level 3 I believe). But this server was not that fast outside United States.

Google in its quest to make the internet faster has stepped into the arena with Google Public DNS. In their own words -

> As web pages become more complex and include more resources from multiple origin domains, clients need to perform multiple DNS lookups to render a single page. The average Internet user performs hundreds of DNS lookups each day, slowing down his or her browsing experience. As the web continues to grow, greater load is placed on existing DNS infrastructure.
>
> Since Google's search engine already crawls the web on a daily basis and in the process resolves and caches DNS information, we wanted to leverage our technology to experiment with new ways of addressing some of the existing DNS challenges around performance and security. We are offering the service to the public in the hope of achieving the following aims:

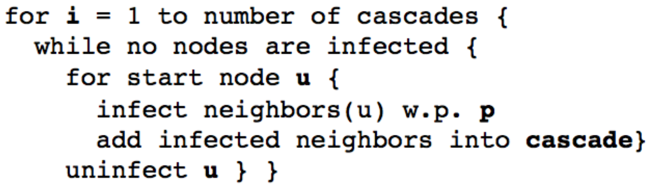
Google Promises us good **Performance, Security and Correct Results**. Correct Results here refer to the practice of redirecting the user to an alternate page when the domain was not found (like OpenDNS does). Google has provided an extensive guide to help users

**Figure 2: Building an Inference Network from Citing in Sources. Here, a except from `manu-j`'s blog shows him citing `en.wikipedia` and `code.google`, inferring the edges from `wikipedia` and `code.google` to `manu-j`**

Finally, on this induced hyperlink graph, we generated a series of cascades following the SIS model. The SIS model is discussed in [6], which largely influenced this project. However, instead of generating the cascades on the post network as Leskovec et al. have done, we generated them on the source network. This was meant to determine whether or not a source is influential across the board regardless of its posts. We expect that high authority nodes would be frequently cited. A cascade model is generated following the algorithm in Figure 3. We intend to study the relative sizes and shapes of cascades to determine what types of conversations people are having on this network.

## 3. MODELING THE INFLUENCE NETWORK

From the memetracker data, we know the number of people who wrote about a certain phrase at a time. For simulation purposes, we divide the time into buckets of 8 hours. Given such $n$ time buckets, we know exactly the number of people who wrote about the phrase during this time bucket. We denote the number of these people by $f(i)$ for the $i^{th}$ time bucket. Our model of influence makes the following



```
for i = 1 to number of cascades {
  while no nodes are infected {
    for start node u {
      infect neighbors(u) w.p. p
      add infected neighbors into cascade}
    uninfect u } }
```

**Figure 3: Building a Cascade Network**

assumptions:

- The probability with which a node $A$ affects its neighbor is $p$; this varies with the time elapsed after $A$ wrote about it.
- The value $p$ is a monotonically decreasing function of the time elapsed.

We expect $p$ to be a monotonically decreasing function because of the effect of *recency* mentioned in [4]. Recency is an effect in the overall news network that favors newer news items over older ones (thereby decreasing the probability of talking about it over time). We try geometric and exponential functions for our analysis:

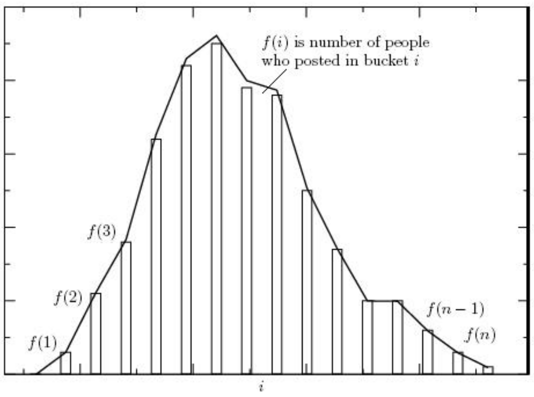


**Figure 4: A diagram depicting $f$.**

$$p(\Delta t) = \frac{p}{2^{\Delta t}}, \qquad (1)$$

$$p(\Delta t) = p^{\Delta t} \qquad (2)$$

Having set the probability, we denote by $N_f(i)$, the number of out-edges from the set $f(i)$. This can be looked as the collective degree of $f(i)$. Thus the influence on a certain level $f(k)$ can be modeled as

$$f(k) = p\,N_f(k-1) + p^2\,N_f(k-2) + \cdots p^{k-1}\,N_f(1) \qquad (3)$$

This holds for $k = 2, 3, \cdots n$. These equations combine to give a easily solvable triangular matrix equation:

$$\begin{bmatrix} p(1) & 0 & \cdots & 0 \\ p(2) & p(1) & \cdots & 0 \\ p(3) & p(2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ p(n-1) & p(n-2) & \cdots & p(1) \end{bmatrix} \begin{bmatrix} N_f(1) \\ N_f(2) \\ N_f(3) \\ \vdots \\ N_f(n-1) \end{bmatrix} = \begin{bmatrix} f(2) \\ f(3) \\ f(4) \\ \vdots \\ f(n) \end{bmatrix}$$

This system can be solved to get $N_f i$s which are then used to make a random graph where we fix the number of out-edges of a level to $N_f(i)$ and each edge between $f(i)$ and $f(j)$ is added with probability $p(|j-i|)$. We repeated the simulations for different values of $p = (0.2, 0.4, 0.6$ and $0.8)$ and we check for the degree distribution in the generated graph. The results for the two models are as follow: We found that the degree distribution is independent of $p$ but it is indeed a power-law with exponent approximately 2 as observed in the hyperlink-network.

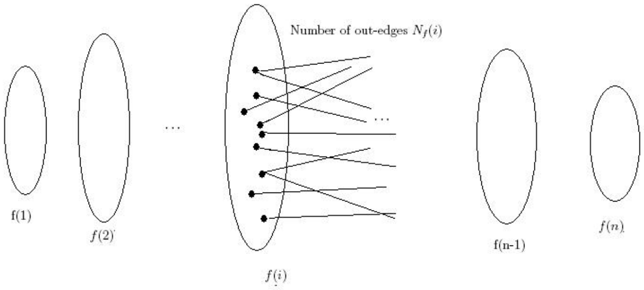


**Figure 5: A diagram depicting $N_f$**

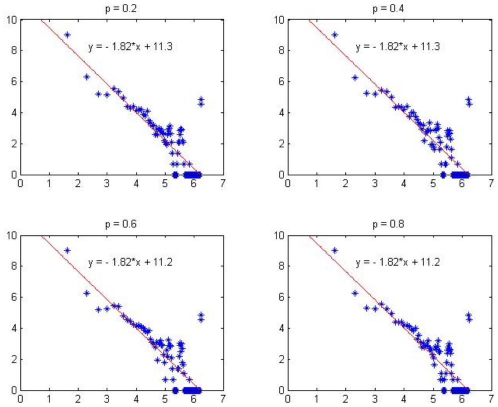


**Figure 6: Log-log plot of degree distribution for geometric decaying p**

## 3.1 Results

### 3.1.1 Frequency Data

By creating and examining the frequency graphs with a 48-hour timestep, we were able to gain some insights into how memes spread in networks. From the t-test, we determined that, on average, blogs produce more posts, discuss topics

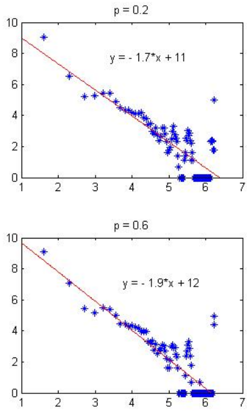


**Figure 7: Log-log plot of degree distribution for exponentially decaying p**

longer, and focus on stories for longer (have a larger diameter) than mainstream media outlets. The peak height is not significantly different. This is not a surprising result, and reinforces the idea that blogs have a significant presence in the news-cycle.

When we separate the memes by category, political data has a significantly larger peak than general data. Considering that the data was collected during the November 2008 election season, it is again not surprising that, when popular, the memes with the largest number of posts were political. Because there were not significantly more posts about political memes, we can conclude that there is somewhat of a "time clustering" effect on politics - more discussion with a common phrase occurs in a tight period of time than in other categories. An interesting result is that the entertainment category has a larger diameter than other categories, refuting the conception that entertainment news is discussed very intensely but only briefly, until the "next big thing" comes along.

When looking at several political memes we categorized as liberal or conservative, we found that conservative topics had more posts per topic and a higher peak (significant at the 7% level). However, we do not know the context in which these quotes were used, so it is difficult to come to a conclusion as to whether conservative topics tend to encourage debate, garner vocal support, or some mixture of the two (so we cannot support or refute the idea that there is a "liberal slant" to today's media). There is no other significant structural difference between liberal and conservative graphs, but we do know that there is more discussion around conservative topics.

Additionally, though there were no significant differences in the slope ratio from category to category, it is interesting to note 91% of the frequency graphs we examined had this ratio less than 1, indicating that the way that a phrase "catches on" and loses focus in such networks has a somewhat predictable pattern.

Finally, we examined the correlations between these statistics in the hope that examining a single feature would strongly indicate how another would behave. Unfortunately, none of the features exhibited a particularly strong correlation (all had magnitudes below 0.7). This result, however, suggests that all the characteristics of the graph we chose to extract are necessary and important for our analysis, as they produced non-trivial results in allowing us to characterize the graphs as we did above.

### *3.1.2 Time Data*

For both generated graphs, we see a power law degree distribution, which is as expected. See Figure 8 for more details.

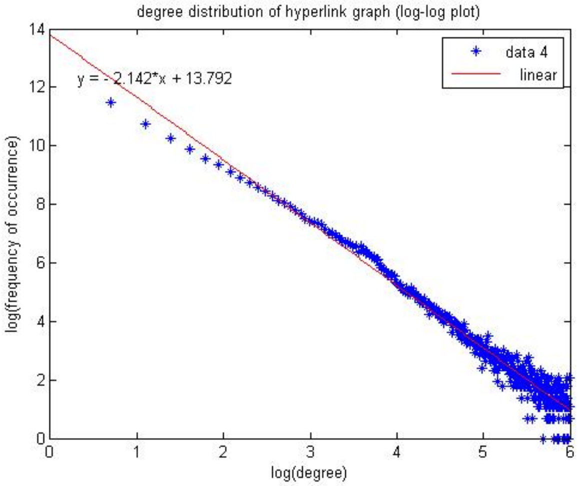


**Figure 8: Power Law distribution on Induced Hyperlink Graph**

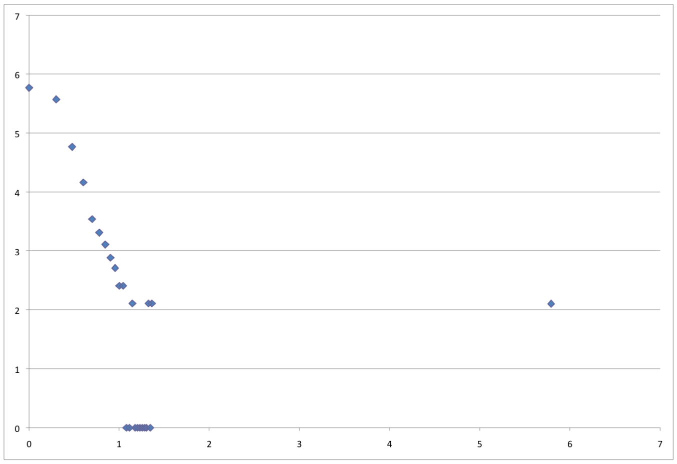


**Figure 9: Log-log plot of connected components size vs. frequency**

There are quite a few large connected components in this graph. There are 128 connected components of size 614,785. Then the next largest connected component is of size 23. The rest of the data follows what we would expect with a large number of connected components of size 1.

The largest connected component includes 97.9% of the graph meaning that the newsmedia is a very well connected industry with very few news sources that do not leverage previously published articles.

Next, on each graph, we try to determine the authority nodes. We determine an authority node by its degree. The phrase graph we generate is undirected so authority is determined simply by the degree. Finding the top authorities in this graph gave us two types of sources: news-media sources (e.g. cnn.com) and official sources for popular events (e.g. Google Android). This is to be expected because people in the overall news network tend to write about current events in either world affairs or business affairs.

The hyperlink influence graph is a directed graph and the number of outgoing links determine how much influence a source has upon other sources. The top authority sources here were a little bit tricky to determine due to the algorithm for parsing urls into sources. Our algorithm separated money.cnn and politics.cnn and health.cnn but it also wanted to separate meghana.blogspot and deepa.blogspot. This allowed us to view which subjects of sources have the most authority, not which sources have the most authority. In this way, we found en.wikipedia and flickr among the top cited (or influencing) sources.

Our cascade results were interesting to look at. Most (98% of) cascades generated were the trivial, one-node cascade. When we look at those cascades with more than one node, we see the following histogram of relative percentage of occurrences (Figure 10).

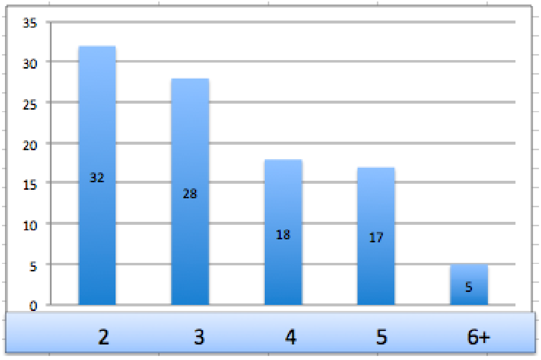


**Figure 10: Percentage of cascades of sizes 2 to 6+**

Of those cascades, we tended to observe two main shapes: a "short and fat tree", or a "tall and skinny" tree. You can see more cascade shapes we observed in Figure 11, taken from [6]

We go through the following analysis of cascades. Cascades with more nodes are considered more "viral", i.e. the disease of meme propagation is more likely to spread. We observed that if we started the cascade algorithm on an authority node (node with high degree), we found that the size of the cascade increased. Additionally, we found that if we choose our root node for the cascade at random and use $p = 2.5\%$ as suggested by the paper, we find very few non-trivial cascades. Increases $p$ increases the size of a cascade, but at a sublinear rate.

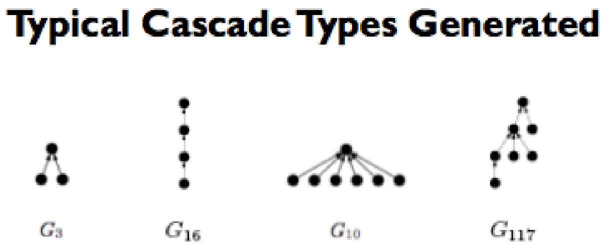


**Figure 11: Some cascade shapes from Leskovec et. al.**

The most typical shape of a non trivial cascade was $G_3$. Conversations that looked like $G_3$ or $G_{16}$ indicate an ongoing conversation, whereas conversations like $G_10$ indicate a very popular news item that might not have much to discuss, but might attract many people. We found that blogs typically had the first type of structure and both news media and blogs had the second type of structure. This relates nicely to the analogous results in our Frequency Data where we found that bloggers tend to talk about a particular phrase for a longer period of time.

### 3.2 Difficulties Encountered

The largest difficulty was memory management. Because we were using Python instead of C++ and because we have very large input files, running the files at once was not possible. In the Time data, we are trying to form a complete graph on every node that shares an phrase $O(n^2)$, which blows the graph up very quickly for large $n$. Additionally, the raw phrase data, which served as the inputs to our hyperlink graph, was so large that running any analysis on it exceeded our computational resources.

## 4. CONCLUSIONS

Overall, we were able to understand a little more about the meme propagation process. The frequency data demonstrated with a high significance many key results among categories and types of news sources propagating news. We found that the inference network and a theoretical model both independently confirmed the same underlying model of determining influence over sources. Additionally, we observed typical types of cascades formed and that the "tall and skinny" cascades (which indicate a long, ongoing conversation) were generated mostly by bloggers. This validated our result from our frequency data stating that bloggers tend to have discussions of a topic for much longer. There are many other future paths to connect our multiple categories of results, which we will describe below.

### 4.1 Future Paths

In the future, by examining a larger number of memes we could obtain a lot more from the frequency data. Because we were limited to hand-categorization, all data except the blog vs mainstream media data was limited to the sample of 200 memes that we hand-classified. Also, there is a portion of the memes that exhibit multiple “humps” and peaks, and it would provide even more information to create a heuristic to deal with these, possibly by finding an appropriate smoothing function (our attempts led to too much reduction of our peak values). Additionally, separating the hyperlink network into the categories we used in the frequency portion and comparing characteristics of the resultant graph would allow us to view differences in the influence of sources on each other in the categories we described.

Based off of the preliminary results from our cascade analysis we have noted a few pathways for future work. One simple idea is to compare the cascades generated with the actual graph cascades to see how similar they are. Another path may be to consider a network that is constructed on posts. In order to construct this network we represent each post as a 3-tuple,

$$P = (T, L, Q)$$

where the variables $T$, $L$, $Q$, refer to time, full hyperlink, and phrase used respectively. We then draw a link from $P_1$ to $P_2$ if the following conditions hold:

1. $T_2 > T_1$
2. There exists some $Q_i$ such that $Q_i \in Q_1$ and $Q_i \in Q_2$
3. There exists some $L_i \in L_2$ such that homepage$(L_i) = D_1$

The first condition ensures that we don’t link to posts in the future. The second condition ensures that if two posts do not have similar phrases, they are not in the same cascade. The third condition ensures that $P_2$ must be linking to an article with $P_1$’s homepage. From this, we can ensure that $L_i$ is the exact address (full url) of $P_1$. This will need to account for the fact that some phrases occur many times on the same homepage but in different articles.

Observing this network can give us more ideas into whether some news sources are more authoritative on certain categories of posts and not authoritative on other categories of posts, or if our assumption that an authority node is an authority node regardless of posts is true.

## 5. ACKNOWLEDGMENTS

The authors would like to thank Mary McGlohon for her help and brainstorming about cascades and Manu Janardhan for allowing us to use his blog in our poster and our paper for demonstration.

## 6. REFERENCES

[1] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. pages 491 – 501, 2004.

[2] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. pages 435 – 443, 2008.

[3] J. Leskovec, L. Backstrom, and J. Kleinberg. Memetracker data. http://memetracker.org/data.html.

[4] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *international conference on Knowledge discovery and data mining,*, pages 497 – 506. ACM SIGKDD, 2009.

[5] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. 2008.

[6] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model. *Society of Applied and Industrial Mathematics: Data Mining*, 2007.

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