Binders Full Of Women: Convergence of Independently-Generated Hashtags on Twitter

CS 224W
Group 48

Diego Pontoriero
dpontori@stanford.edu

Evelyn Gillie
egillie@stanford.edu

Abstract

In social networks, topic propagation is an area of obvious interest and frequent study. However, the methods by which topics arise in the first place is less understood. Though some topics may arise from purely internal memes, many are derived in real-time from events affecting users in the real world.

We analyze a large sample of Twitter data with a focus on topic (hashtag) creation as a response to external stimuli. We present several examples in which disparate topic instances converge into a small prevailing set. We also propose a simple simulation model that reasonably matches observed behavior.

1 Introduction

Twitter and other social media sites provide a unique view into the real-time zeitgeist: by being constantly connected to social networks via laptops and smartphones, users have effectively transformed news *consumption* into news *participation*. Accordingly, social media sources are now a fixture in media reporting, since they provide an instantaneous and interactive perspective of what millions of internet users are interested in and passionate about. The combined effect of the 2012 United States Presidential election, 2012 Summer Olympics, and other significant, globally-visible news events has resulted in a landmark year for Twitter activity. Twitter records were broken [1] and re-broken [2], [3] with each successive event, and some topics even took on internet lives of their own. For example, the phrase "binders full of women," fatefully uttered by Governor Mitt Romney during the second Presidential Debate resulted in a flurry of responses on Twitter, Facebook, and Tumblr [4].

In order to detect trends from thousands of Twitter posts, Twitter algorithmically extracts *trending topics* based on the content of the tweets. Additionally, an organically-developed but de facto-standardized mechanism called a *hashtag*—a word or phrase preceded by a `#` symbol, such as `#bieberfever` or `#yolo`—is employed by Twitter users to explicitly mark a tweet as belonging to a certain topic.

Several previous works [5], [6], [7] address the question of how topics propagate throughout Twitter, i.e. using contagion models. However, the mechanism by which real-world events actually transform into specific Twitter-wide memes in the first place remains an open question. In this paper we investigate this mechanism by aligning large samples of Twitter data with instances of significant external events and searching for criteria that cause certain memes to arise and, in some cases, subsequently flourish.

Analyzing and modeling convergence of independently-generated hashtags on Twitter. Our analysis is concerned with Twitter topics from the time of their inception to the time that they become trending, i.e. are being utilized by a relatively large number of Twitter users. Intuitively, in order for a new topic to flourish it must be relatable by a large number of people. Accordingly, we have selected several examples of widely-watched events (Olympic events, Presidential Debates) where notable external stimula, such as memorable phrases or occurrences, were directly responsible for topics that became trending on Twitter. Using a large sample of Twitter data we are then able to identify topics that were created as a response to this stimulus and track them to see which flourish and which do not.

A drawback to our approach is that for financial and logistical reasons we do not have access to the entire Twitter social graph or full stream of Tweets. Therefore, in addition to a qualitative and quantitative analysis of Twitter data, we also present a simulation model that provides additional insight into the observed results. This model utilizes a simple scale-free network [8] to simulate the propagation competing meme instances.

The remainder of this paper is organized as follows. First, we discuss relevant prior work in the area of social network analysis and meme propagation. Second, we outline the method by which we assembled our data set and provide some high-level descriptive statistics. Third, we describe our analytical approach of the data and our initial findings. Fourth, we introduce a simulation model for meme creation and propagation. Finally, we conclude and discuss areas for further work.

2 Related Work

Due to its size, data availability, and certain interesting properties, Twitter is a prevalent subject of study in the literature. Kwak et al [10] present an empirical overview of the Twitter social graph structure and user behavior. The authors observe that the graph generally follows a

power law degree distribution with a few small deviations due to initially suggested people to follow, a restriction on the maximum number of people to follow, and an unexplained long tail which the authors believe is due to a preponderance of celebrities.

Several prior works directly address the topic of Twitter hashtag propagation. Romero et al [6] present a broad empirical study of hashtag propagation on Twitter, parameterize a model based on their observations, and then run a simulation based on these parameters. By hand-classifying hashtags into a number of qualitative categories the authors uncover several differences in the propagation of hashtags as a function of their categories. However, the authors are primarily concerned with measures of stickiness and persistence of already-established hashtags and do not address their formation.

Myers and Leskovec [7] propose a model that extends the Independent Cascade Model to account for the effect of prior exposures to infections on a the probability of acquiring new infection. The motivation for the model is based on the insight that information does not spread in isolation, but rather many contagions propagate through a network simultaneously and either cooperating or competing with one other for survival. They formulate a statistical model that uses three factors—inherent content virality, inherent user bias to share, and a content interaction term that captures the cross-content effect of exposure to one piece of information has on the adoption of others—to compute the probability that a user will adopt a new piece of information. However, whereas the authors are only concerned with users analyzing their Twitter feeds and deciding which items to re-share, we are interested in the origins of these memes as well as their subsequent adoption.

Leskovec et al [5] examine the effect of external sources on information propagation through Twitter by measuring the lag in peaks of attention from news outlets to blogs and online media. Their work is primarily concerned with developing an algorithm for generating variants of phrases so that they might increase recall of discussion about a given phrase. Compared to our work, their approach only handles information flowing from a single authoritative source, e.g. the person quoted. Our work, on the other hand, considers the interpretation of real-time primary sources, e.g. live television, by many internet users who then go on to share and consume these interpretations online via social media.

Ratkiewicz et al [11] present a framework for detecting and tracking memes in relatively real-time as they arise in Twitter. This system uses a hand-curated list of 2500 keywords to filter and extract politically-oriented tweets. The purpose of this system is to perform diffusion and sentiment analysis in order to detect abusive or malicious behaviors. Though the system utilizes hashtag coincidence to associate new hashtags with known political ones, it does not explicitly study the creation and convergence of hashtags for a particular meme.

3 Real-World Meme Creation

3.1 Data Collection

In August, we started collected Twitter data using Twitter's streaming API on an EC2 instance. This collected around 55-60 tweets per second during peak hours, and 25-30 during non-peak hours. For the third Presidential debate and election day, we filtered the stream on the keywords 'obama' and 'romney' in order to get a more precise stream on memes related to the debate and election, and after the election, we returns to collecting the random stream.

3.2 Data Processing

Since we were only interested in hashtags, we parsed the tweets and indexed the hashtags by 10-minute intervals. We did not think it necessary to perform any stemming, but did lower-case all tags.

3.3 Analysis

We focus on notable events with significant twitter activity, and measure the popularity of unique hashtags representing a common idea over time. Three events that we inspect are 1) the anniversary of the 9/11 attack, 2) the third presidential debate, and 3) the November 6 general election. Graphs of hash tag frequencies by hour are shown in Figure 3 and Figure 2. These graphs illustrate two examples of hashtag convergence over time; observe the set of 9/11 hashtags: at 9am PST there are several competing hashtags describing 9/11, of which none has more than sixty occurrences, and only five have more than twenty occurrences: `#neverforget`, `#9/11`, `#11september`, `#september11th`, and `#911`. By noon PST, #neverforget is by far the most popular tag with 468 uses per hour in our stream. The next most frequent hashtag is `#remember911` with 188 occurrences; note that `#remember911` was not in our top five at 9am. By 4pm `#remember911` is outperforming `#neverforget`, and ends the day at the top.

Similarly, election day starts with many variations of 'vote' hashtags: `#votedem`, `#voteobama2012`, `#vote4obama`, `#voteobama12p`, etc., but users flock to `#voteobama`. As an interesting side note: `#voteforromney` was the most popular romney-related hashtag until `#voteobama` became popular, at which point `#voteromney` became more popular, perhaps as a response.

We look at a few measures to detect predictors of hashtag success: the influence (in terms of total number of followers) of users using a particular hashtag (we call this "impressions", and what proportion of hashtag usages were in retweets. The intuition behind measuring retweets was to examine if there was some sort of innate property of a hashtag that lent itself to retweeting. In some instances, impressions appeared to be a solid indicator (8). The proportion of tweets being retweets was noisy and did not yield any meaningful results.

In all, we notice an early jumble of seemingly independently-created hashtags, to which users flock to one or two. Which leads us to the question of modeling: how might we model this evolution of hashtag selection, and what does it say about how Twitter users choose to adopt hashtags?

4 Modeling Meme Creation

In this section we describe a simulation model that demonstrates how our observed results can arise in a real-world graph. We present the model formulation in intuitive and formal terms, discuss our rationale for choosing each parameter, evaluate the performance of our model, and discuss implementation choices and tradeoffs.

4.1 Motivation

The intuition behind our model is based on a simple representation of a typical Twitter user responding to a significant, specific news event. In our model, at some point during the course of a given day the user logs in to Twitter, reads the most recent tweets of the users that he follows, and then decides to tweet something himself.

His decision to tweet is based both on dominating external stimuli and the tweets that he just observed. If it is election day, for example, a user may wish to express his vote for Barack Obama using a hashtag. On the one hand, if he did not see any tweets containing hashtag related to this topic, he may invent his own in an effectively random manner using relevant keywords, perhaps arriving at `#vote4obama`. On the other hand, if he observes several tweets in his feed that use the hashtag `#obama2012` it would be more natural for him to "join the conversation" and include that hashtag instead.

Rather than attempting to simulate the entire Twitterverse and all possible memes and hashtags, for simulation purposes we consider only the subset of Twitter that is concerned with a single topic (such as election day in the example above). This simplification makes the model self-contained and allows a more natural and direct comparison with the data we were able to collect (Section 3.1).

4.2 Formal Description

A high-level view of the model is described in Algorithmx 1, and detailed presentation follows. Our model has the following parameters:

- $G(n, m)$: A directed graph that represents the Twitter social graph structure. A node u is said to follow a node v if there exists a directed edge (u, v). Thus the number of followers for a node is equal to its in-degree.
- L: A list of words that constitute a potential hashtag amalgamations. Examples of word lists include $\{vote, for, obama, 2012\}$ and $\{remember, september, 911\}$.
- $f(t)$: A function which maps each time step t to the fraction of the nodes in G that become active at t. For example, at a given time step 5% of nodes may decide to tweet.

We begin by computing $P(L)$, which is defined as the list of all permutations of all subsets of the words. This is used as a proxy for different hashtags that might arise in response to a live event. Illustrative examples of different instances of $P(L)$ are shown in Figure 4.

The main phase of the simulation now occurs. At each time step t we choose a fraction $f(t)$ of the nodes in the graph uniformly at random without replacement. For each node n we then choose one hashtag from $P(L)$ to tweet. The probability of n choosing a particular hashtag is initially uniform across all hashtags in $P(L)$. However, before n chooses a hashtag we iterate through all the nodes that n is following, i.e. has directed edges to, and compute a histogram of the number of hashtags that n sees those nodes tweet. This histogram is used to re-weight the probability distribution (the specific mechanism is described below). Then n chooses a hashtag according to the updated probability distribution and "tweets."

The simulation repeats for each time step in the domain of f, with nodes observing the tweets of their neighbors from previous rounds. Note that though nodes are chosen without replacement in each round, they are replaced between rounds. Accordingly, in a subsequent time step the a particular node may tweet again, and even choose a different hashtag to tweet.

The current model has two outputs: (1) a plot of the number of tweets of each hashtag at each time step, and (2) a plot of the number of impressions of each hashtag at each time step, where impressions is defined as the number of tweets for a hashtag multiplied by the number of followers of the nodes who tweeted it.

4.3 Parameter Settings

4.3.1 G: Social Graph

Due to API limitations we were not able to access the complete Twitter social graph. Instead, in our simulation we create G using a randomly-generated scale-free directed graph, which we generate using the method described by Bollobas et al [8] and implemented by `NetworkX` [9]. Though Kwak et al [10] show that the Twitter social network does not follow a power law degree distribution perfectly (they ascribe the higher-than-expected number of highly-followed nodes to the presence of many celebrities), we believe that it is sufficient for our simulation. In fact, as discussed below, we think that a higher proportion of celebrities would only increase the likelihood of converging on a single hashtag.

For our simulations we generate G with 1000 nodes, utilizing the default `NetworkX` parameters of $\alpha = 0.41, \beta = 0.54, \gamma = 0.05, \delta_{in} = 0.2$, and $\delta_{out} = 0$ to generate the graph.

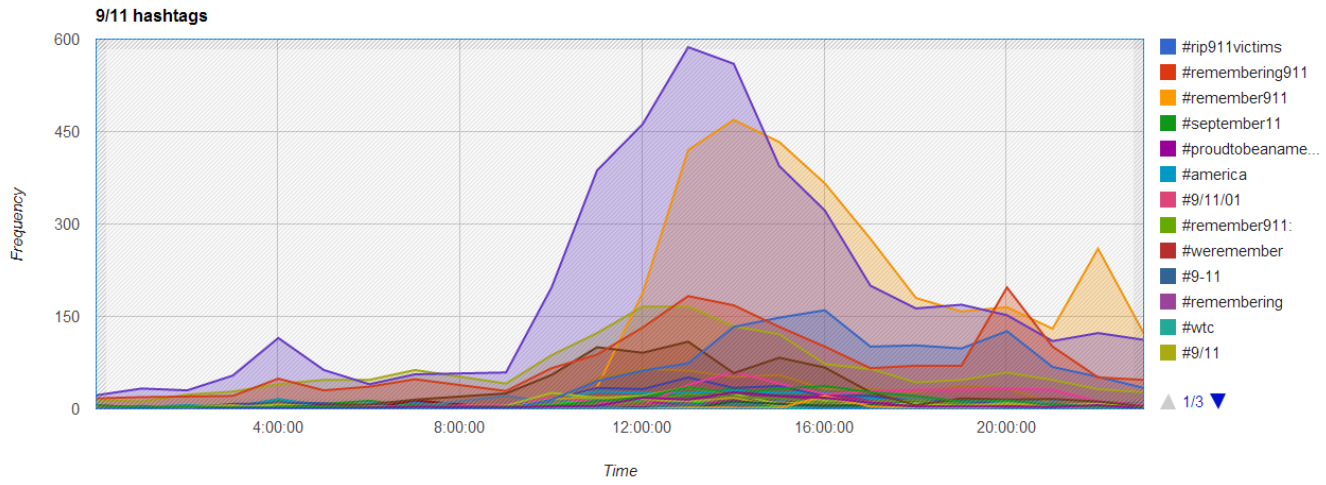


Figure 1: Frequencies of 9/11-related hashtags by hour.

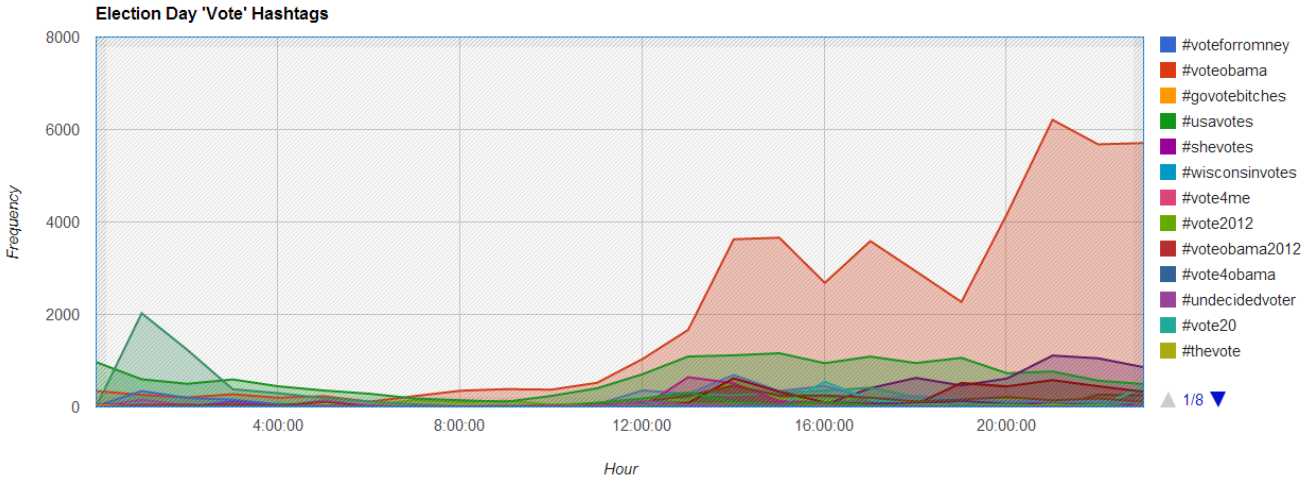


Figure 2: Frequency of voting-related hashtags by hour.

4.3.2 L: Word List

The length and content of the word list does not significantly affect the simulation, as long as the ratio between $P(L)$ and the number of tweets per time step is within 1-2 orders of magnitude. For our simulation we used the $L = \{horses, and, bayonets\}$, which results in 15 permutations.

4.3.3 $f(t)$: Fraction of Hashtag Tweets

We used real-world data as the basis of $f(t)$ for our simulations. Our simplifying assumption of treating our graph as the subset of Twitter concerned with a particular topic (such as election day) allows us to derive $f(t)$ empirically: for a given topic and time step t, compute the fraction of users that tweeted a hashtag related to the topic out of all tweets related to the topic.

For example, Figure 5a shows the fraction of tweets containing a voting-related hashtag out of all voting-related tweets over the course of election day. The values of $f(t)$ used in our simulation are based on this data. Other topics exhibited similar ratios. As we show in our discussion of parameter sensitivity below, $f(t)$ has a noticeable impact on results.

4.4 Evaluation and Discussion

A trace of a single trial of the simulation using the parameters described in the preceding section are shown in Figure **??**. As the figures illustrate, the simulation converges on a single hashtag (the fictitious and nonsensical tag `#and`). More importantly the trace matches the observered real-world election day hashtag propagation behavior!

We define convergence explicitly as follows: for each hashtag in $P(L)$ we compute the maximum number of tweets that occurred at any time step. We then compute the mean and standard deviation for all hashtags. If the maximum number of tweets for at least one hashtag

Since the graph is generated randomly, convergence is not guaranteed for all invocations of the simulation. Instead, we can run a Monte Carlo-esque simulation and

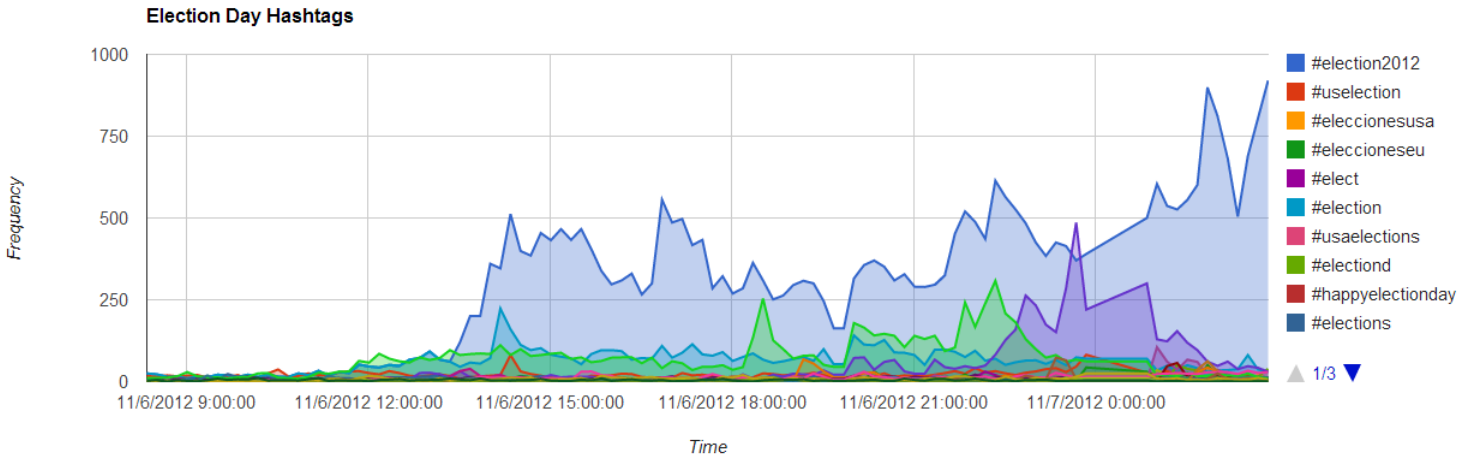


Figure 3: Frequencies of Election hashtags by 10-minute intervals.

Words	Hashtags
{*kill, big, bird*}	`#bird #birdbig #big #birdkill #killbird #birdbigkill #birdkillbig #bigbirdkill #bigkillbird #killbirdbig #killbigbird #bigkill #killbig #kill`
{*bunch, of, malarkey*}	`#malarkey #malarkeyof #ofmalarkey #of #malarkeybunch #bunch #bunchmalarkey #malarkeyofbunch #malarkeybunchof #ofmalarkeybunch #ofbunchmalarkey #bunchmalarkeyof #bunchofmalarkey #ofbunch #bunchof`

Figure 4: Examples of generated meme permutations.

compute the fraction of invocations that converge. For the given parameters and $n = 50$ trials we observed a convergence rate of 86%.

In developing and testing the model we noticed that changing the number of tweets had a visible impact on convergence probability. Figure 5b shows a sensitivity analysis of scaling our empirical $f(t)$. A larger multiplier results in more nodes being chosen at random to tweet at each time step. It appears that doubling the fraction of nodes that tweet hashtags at each time step is enough to nearly guarantee convergence of a single hashtag. The real-world values are less stable, which matches our experience with some topics (3).

In general, our model matched the observed behavior of hashtag propogation. 8 shows an example where our impressions graph roughly matched a simulation impressions graph, and the resulting hashtag propgation looks similar. In this example, a highly-followed nodes uses a hashtag, resulting in a spike in impressions, and the hashtag evolutions propogate very similarly. Simulations without an early "injection" of impressions evolve similarly to the real-world graphs shown in 3 and 2.

4.5 Alternative Model Formulations

Though we believe that our model has both an intuitive interpretation and quantitative validation, we submit that it is not as rigorously motivated or quantitatively validated as we would like. However, there are a few areas in addition to the sensitivity analysis described above where we were able to explore alternative formulations for comparison.

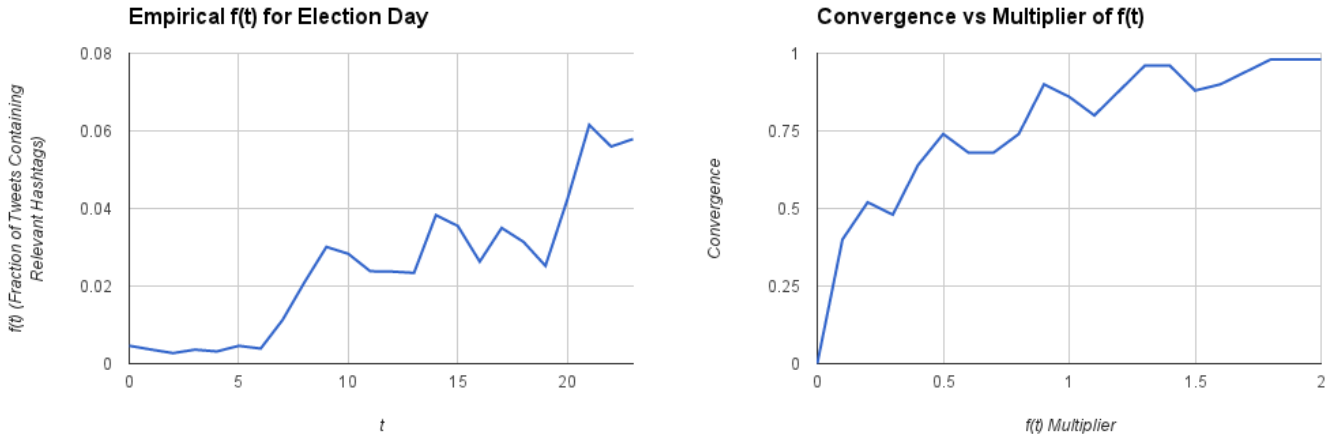
4.5.1 Re-weighting Function

One area of the model that allows for different formulations is in the mechanism by which a node updates the probability of choosing a particular hashtag in response to observing the tweets of neighbors. Ideally we would like to calibrate this on real-world data. Unfortunately, as described in Section 3.1 we could not think of a way to directly observe a user's decision-making process when selecting a particular hashtag to adopt.

Instead, we considered three possible hypotheses and implemented three different kernels that attempt to behave in accordance to those hypotheses:

1. **Equal Weight**: In this model we treat the influence of all tweets equally. The intuition behind this model is that when looking at his tweet feed a user does not discriminate among sources of tweets, only content.

2. **Celebrity Influence**: In this model we weight the influence of a tweet by the influence of the tweeter, which we define as the number of followers (in-degree) that the tweeter has. The intuition behind this model is that a user is more likely to replicate/retweet the



(a) Empirical $f(t)$ for election day tweets.

(b) Convergence vs. scaling $f(t)$ by a multiplier.

Figure 5: Empirical $f(t)$ and convergence sensitivity.

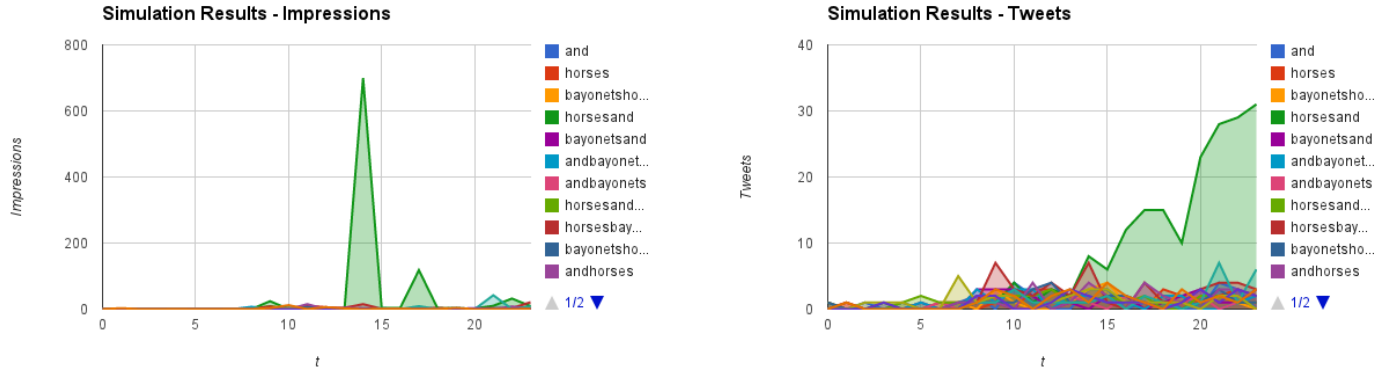


Figure 6: An example simulation

behavior of someone who is seen as having stature or authority.

3. **Community Influence**: In this model we weight the influence of a tweet inversely to the influence of the tweeter. The intuition behind this model is that a user is more likely to replicate the behavior of his close acquaintances. Ideally we would use community detection, i.e. betweenness, to model this but as a simplifying proxy we observe that due to the power law, for the vast majority of users their close friends will not have a high in-degree.

Figure 7a compares the results of using the three different kernels. The Celebrity Influence kernel performs the best, whereas the Community Influence kernel performs very poorly. This makes some sense, since a meme will have a greater chance of becoming dominant if many people are exposed to it *and* choose to adopt it.

In addition to the three kernels described above, an additional parameter k controls the effect of observed tweets on the null probability of choosing a hashtag uniformly at random. Figure 7b presents a sensitivity analysis of k for each kernel. Lower k means that observed tweets have a lower effect on the probability distribution ($k = 0$ means no effect). There effect of varying k is not very pronounced, though it appears to be slightly positive. Again, this makes some intuition, since a very high k has the effect of skewing the probability distribution to give the node effectively little choice on which hashtag it will adopt.

5 Conclusions and Further Work

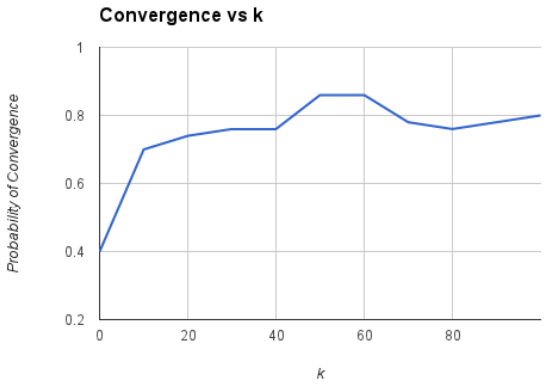
Though there has been a significant amount of research on the propagation of memes using Twitter as a medium, very little has been done to analyze the process by which external influences cause memes to emerge in various forms before coalescing into single, dominating instances. In this paper we have presented two key contributions. First, we have analyzed Twitter data and uncovered several empir-

Set $P(L)$ to all permutations of L
for $t \leftarrow 1$ **to** T **do**
  Select $M = nf(t)$ nodes uniformly at random from $G(n, m)$
  for node $\in M$ **do**
    Set probabilities $p(h_i)$ of $P(L) = h_1 \dots h_k$ hashtags to $\frac{1}{|P(L)|}$
    for succ $\in$ *successors(*node*)* **do**
      if succ *tweeted hashtag* h' **then** update probability $p(h')$
    end
    Randomly choose hashtag from $P(L)$ weighted by p and tweet
  end
end

Algorithm 1: Hashtag Population Simulation Model

Kernel	Convergence Probability
Equal Weight	0.90
Celebrity Influence	0.96
Community Influence	0.46

(a) Convergence probability for three different re-weighting kernels.



(b) Convergence probability as a function of k.

Figure 7: Re-weighting sensitivity analyses.

ical examples of meme emergence and coalescence in response to external stimuli. Second, we have formulated a simple yet intuitive and explanatory model that matches the observed real-world behavior.

There are many areas which, if given more time, we would have liked to refine or explore further. First, if possible we would like to access the real Twitter social graph. Additionally, we would like to investigate modifying the generated scale-free graph to create graphs that more closely match Kwak et al's observed characteristics. Additionally, for our "Community Influence" re-weighting kernel we would have liked to compute betweenness and derive true communities for weighting influence.

There are some more ambitious goals which were outside our original scope but still seem to be potentially interesting avenues to investigate. For example, our model does not include any linguistic analysis, which may be interesting for refining the hashtag formulation function.

References

[1] Olympic (and Twitter) Records. *Twitter Corporate Blog*, August 12, 2012.

[2] Camia, Catalina. Presidential debate sets record on Twitter. *USA Today*, October 4, 2012.

[3] Vena, Jocelyn. President Obama's Historic Photo Steals Twitter Record From Justin Bieber. *MTV NEWS*, November 7, 2012.

[4] Chai, Barbara. 'Binders Full of Women' Spawns Three-Ring Circus on Web. *Wall Street Journal Speakeasy Blog*, October 17, 2012.

[5] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. *KDD'09*, 2009.

[6] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. *WWW 2011*, 2011.

[7] S. A. Myers and J. Leskovec. Clash of the Contagions: Cooperation and Competition in Information Diffusion. 2012.

[8] B. Bollobas, C. Borgs, J. Chayes, and O. Riordan. Directed Scale-Free Graphs. *Proc. 14th ACM-SIAM Symposium on Discrete Algorithms*, 132-139, 2003.

[9] NetworkX. `http://networkx.lanl.gov/`

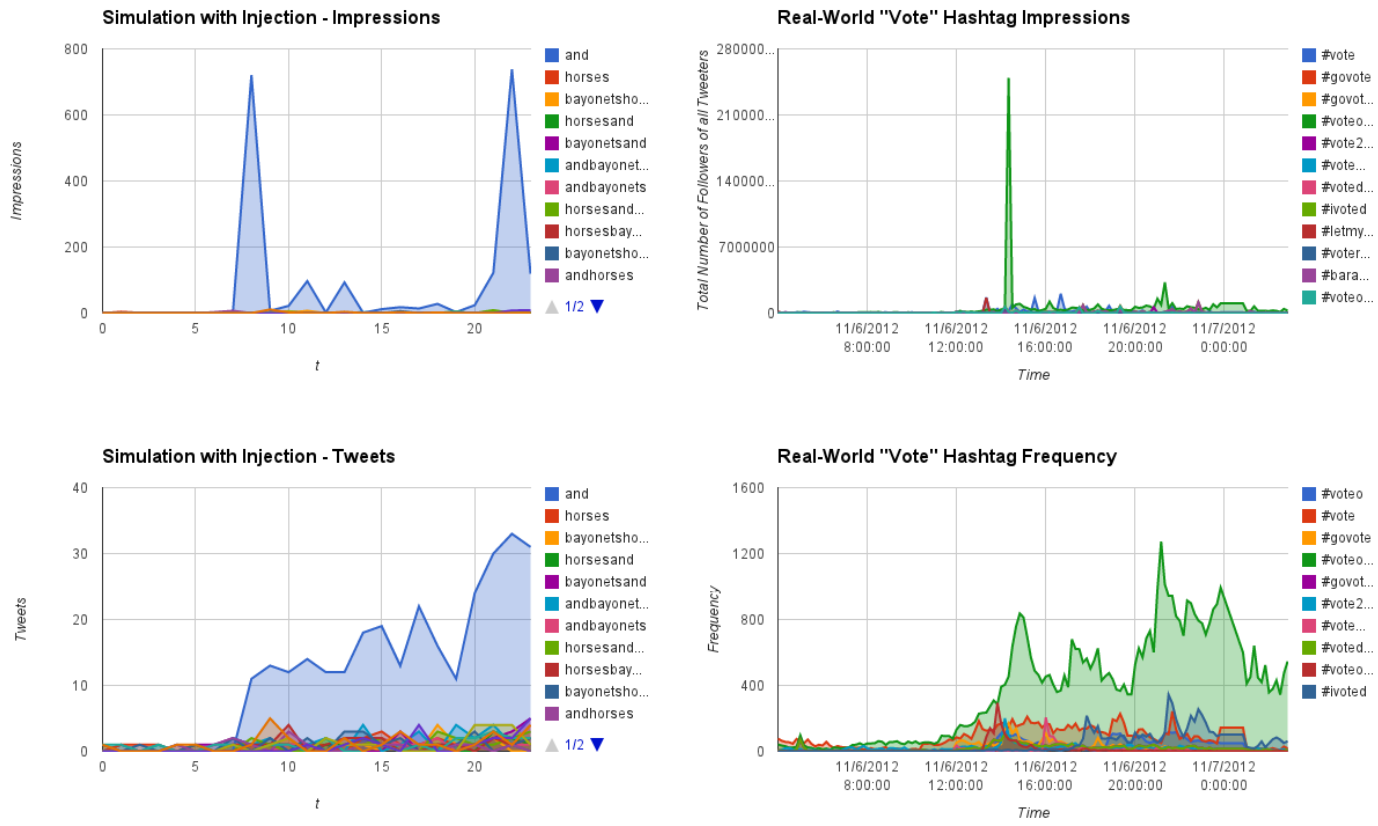


Figure 8: Comparsion of Impressions and Hashtag Usage in our Simulation and Real-World Twitter stream

[10] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? *WWW 2012*, 2010.

[11] J. Ratkiewicz, M. Conover, M. Meiss, B. Goncalves, S. Patil, A. Flammini, F. Menczer. Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams. *WWW 2011*, 2011.