Temporal Analysis of Information Cascades on Twitter

Maryam Aly

May 2012

Advisors:
Luis von Ahn
Brendan Meeder

Carnegie Mellon University
Completed as part of my Senior Thesis project.

Abstract

We are analyzing how a piece of information spreads on a global scale. Specifically, we are analyzing how the centrality of the spreading network changes over time, as well as how separate components merge within this network. We are also exploring the differences between how pieces of information endogenous to a network and pieces of information exogenous to a network spread.

The data that we are using to study these phenomena is from Twitter. We are examining information in the form of hashtags on Twitter, and the underlying network that we are using is the direct message network.

Contents

1 Twitter

Twitter is a site which allows users to broadcast short messages (140 characters or less) to the site. A tweet is one of these short messages. An @-message is a tweet which contains the @ character followed by the Twitter handle of a user. A hashtag is a # symbol followed by a word, phrase, or an abbreviation that refers to some idea relevant to the tweet.

The Twitter data set that I am working with was gathered over the time from August 2009 to January 2010. For each user whose Twitter userid was less than 10 million, each user's 3,200 most recent tweets were collected as well as all the friends and followers of each user. So, if the user tweeted fewer than 3201 tweets since the account was created, then all of the users tweets were recorded. The friends and followers of these users were crawled in the same way. This data set includes about three billion tweets from over 60 million users.

2 Previous Work

A large amount of work has been performed in the space of social networks. Much of this work has been on information cascades and the structure of the networks.

There have been studies on stickiness– how many times a person must see a piece of information before he or she will spread it to others. There is also a notion of persistence– how long a piece of information will remain in the network. Both of these properties have been studied within the Twitter social network under the context of topical analysis [4]. With high accuracy, one can predict whether there will be an edge between two users [3]. One can predict whether an edge, once we know that it exists, is a positive or negative edge [9].

Once triadic closure is achieved, there is this notion of balance– since all three nodes form a clique, more information will pass between any two pairs of the triangle– and a notion of exchange– since the pair that just became closed now have a direct link, the communication between this pair and the third node will decrease. Both of these phenomena are documented in the social sciences, but they are also present within the Twitter network [6].

One study found the cost of having a differed opinion from your neighbors in a social network [2], and another study was able to predict if a network will split based on the friendliness between users of different opinions [5].

The cascade types of blogs are indicative of what community the blog belongs in. The number of cascades is an indicator to the types of cascades that were present in the blog development [12].

When choosing to join a social network, the number of friends that a person has, in addition to the way that those friends are connected are very important to the decision of the choice [11]. Even though Twitter does not supply the creation date of an account, it can be predicted with high accuracy using the times that a particular account followed celebrity Twitter users [3].

One study found that information that pertained to bad news had a shorter

lifespan within the Twitter network than information that pertained to good news. The information studied in this case was urls instead of hashtags [7].

When it comes to maximizing the spread of influence, we have a greedy algorithm which performs at an accuracy of above 63% [8].

One study focused on the formation of large, connected, real-world networks. It identified all of the components as the graph evolved over time, and found a model for the rebel probability, or the probability that any given component avoids becoming connected to the largest connected component. It was shown that the rebel probability decreased exponentially over time [14].

3 Results

Last semester, my work revolved around finding a model of centrality by using solely the Twitter @mention graph data of the first million users. I was looking at certain case-study hashtags. These hashtags were among the most popular during the time that the data was collected and included:

- #mw2– the video game Call of Duty: Modern Warfare 2, a first-person shooter which was released during the time of the Twitter data and went on to win multiple awards
- #ff– the hashtag which represents Follow Fridays on Twitter. Every Friday, Twitter users suggest other users who are worth following with this hashtag.
- #tcot– the hashtag which represents Top Conservatives on Twitter. Twitter users suggest other users who are conservative and worth following with this hashtag.
- #mj– Michael Jackson, who died during the time of the data collection.
- #bbc– British Broadcasting Corporation

In order to obtain the centrality of the spread of these hashtags, I extracted out the subgraph network for each hashtag from the original dataset and the timestamp for each edge. Note that an edge occurs from user 1 to user 2 if there is an edge from user 2 to user 1 in the @mention graph and user 2 tweeted the hashtag after user 1.

Using this data, I would calculate the centrality over time, recalculating approximately every 6 days. The centrality measure that I used is graph closeness centrality which is defined by

$$\frac{\sum_i [C_c(v^*) - C_c(v_i)]}{(n-1)(n-2)/(2n-3)} \text{ where } C_c(v_i) = \frac{n-1}{\sum_{j \neq i} d(v_i, v_j)} \text{ and } C_c(v^*) = \max_i C_c(v_i).$$

Through this, I had no conclusive results as timeframe of the data set put limitations on the comprehensiveness of the centrality graphs for specific hashtags. Specifically, there was no way to verify if the centrality curve for the data

was the entire curve, or just a piece of the entire curve constrained to a smaller timeline.

Moving forward, we want to determine whether a theoretical approximate of an endogenous spread is different from that of an exogenous spread by running simulations on the following different models of graphs and studying their centrality over time:

- Watts-Strogatz model
- Small world model
- Erdos-Renyi model
- Barabasi-Albert model
- underlying @-mention graph used already

3.0.1 Watts-Strogatz

This is a small-diameter graph. To construct this graph, you start with a regular graph, then pick a constant k. Now, every node becomes adjacent to k more nodes, chosen uniformly at random.

3.0.2 Small World Graph

This is also a small-diameter graph. It has the added guarantee that local routing is efficient; if you want to route a message from node a to node b, then you can do so by routing through neighbors in $\Theta(\log n)$ steps where n is the total number of nodes in the graph. To construct this graph, you start with a regular graph, then pick a constant k. Now, every node becomes adjacent to k more nodes, but the probability that edge (i, j) forms is proportional to $1/d(i, j)^2$.

3.0.3 Erdos-Reyni

Each potential edge has equal probability of appearing in the construction of this graph. If you want a graph on n nodes with m edges, then the probability that edge (i, j) forms is

$$\frac{m}{\binom{2n}{2}}.$$

3.0.4 Barabasi-Albert

This is a power-law graph. We grow the graph by starting with a connected graph, and with each new node, connecting the node to k other nodes with probability proportional to the degrees of the nodes. This creates nodes with very high degrees, which will be similar to the celebrities on Twitter.

3.0.5 Twitter Network

We can still use the @-mention graph, but with simulated cascades instead of actual cascades.

References

[1] Easley, David, and Jon Kleinberg.*Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. New York: Cambridge UP, 2010. Print.

[2] D. Bindel, J. Kleinberg, S. Oren. How Bad is Forming Your Own Opinion? Proc. 52nd IEEE Symposium on Foundations of Computer Science, 2011.

[3] J. Cheng, D. Romero, B. Meeder, J. Kleinberg. Predicting Reciprocity in Social Networks. Proc. 3rd IEEE Conference on Social Computing, 2011.

[4] D. Romero, B. Meeder, J. Kleinberg. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. Proc. 20th International World Wide Web Conference, 2011.

[5] S. Marvel, J. Kleinberg, R. Kleinberg, S. Strogatz. Continuous-Time Model of Structural Balance. Proc. National Academy of Sciences, 108(5) 1771-1776, 1 February 2011.

[6] D. Romero, B. Meeder, V. Barash, J. Kleinberg. Maintaining Ties on Social Media Sites: The Competing Effects of Balance, Exchange, and Betweenness. Proc. 5th International AAAI Conference on Weblogs and Social Media, 2011.

[7] S. Wu, C. Tan, J. Kleinberg, M. Macy. Does Bad News Go Away Faster? Proc. 5th International AAAI Conference on Weblogs and Social Media, 2011.

[8] D. Kempe, J. Kleinberg, E. Tardos. Maximizing the Spread of Influence through a Social Network. Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2003.

[9] J. Leskovec, D. Huttenlocher, J. Kleinberg. Predicting Positive and Negative Links in Online Social Networks. Proc. 19th International World Wide Web Conference, 2010.

[10] J. Leskovec, L. Backstrom, J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2009.

[11] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. Proc. 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2006.

[12] M. McGlohon, J. Leskovec, C. Faloutsos, N. Glance, and M. Hurst. Finding patterns in blog shapes and blog evolution. International Conference on Weblogs and Social Media. Boulder, Colo., March 2007.

[13] U. Kang, M. McGlohon, L. Akoglu, and C. Faloutsos. Patterns on the Connected Components of Terabyte-Scale Graphs. IEEE International Conference on Data Mining (ICDM10). Sydney, Australia, December 2010.