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 Introduction

We frequently make decisions based on what people decided before us. For example, you probably do not own an HD DVD Player, but you might own a Blu-ray Player. Imagine you were in the market for a new media player, and you are trying to decide between an HD DVD Player and a Blu-ray Player. You might decide that an HD DVD Player is superior than a Blu-ray Player. However, before you buy, you observe that many of your friends have already bought Blu-ray Players. When you actually make your purchase, you decide to buy a Blu-ray Player instead of the HD DVD Player. An information cascade has occurred.

An information cascade happens when agents are making a decision with a small number choices sequentially, and the agents base their decisions rationally on what others decided before, independent of their personal knowledge.

One real-world phenomenon that is modeled as an information cascade is hashtagging on Twitter.

2 Twitter

Twitter is a website which allows users to broadcast short messages (140 characters or less) to the site. A tweet is one of these short messages.

2.1 @-Messages

An @-message is a tweet which contains the @ character followed by the Twitter handle of a user. When user A @-mentions user B, user A is trying to get the attention of user B. Therefore, we can think of an @-mention as signaling a influence relationship– user B influences user A.

2.2 Hashtags

A hashtag is a # symbol followed by a word, phrase, or an abbreviation that refers to some idea relevant to the tweet.

3 Previous Work

In "Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter," Romero et al found that we can describe the cascade of a particular hashtag based on two properties– stickiness and persistence. Stickiness describes the number of exposures required before a user makes the decision to use a particular hashtag.

Persistence describes the marginal effect that each new exposure has on the exposure chances of a particular user [4].

In Figure 1, we have an example exposure curve, which tells us the probability that a user will use a particular hashtag after k exposures, but before the k+1st exposure.

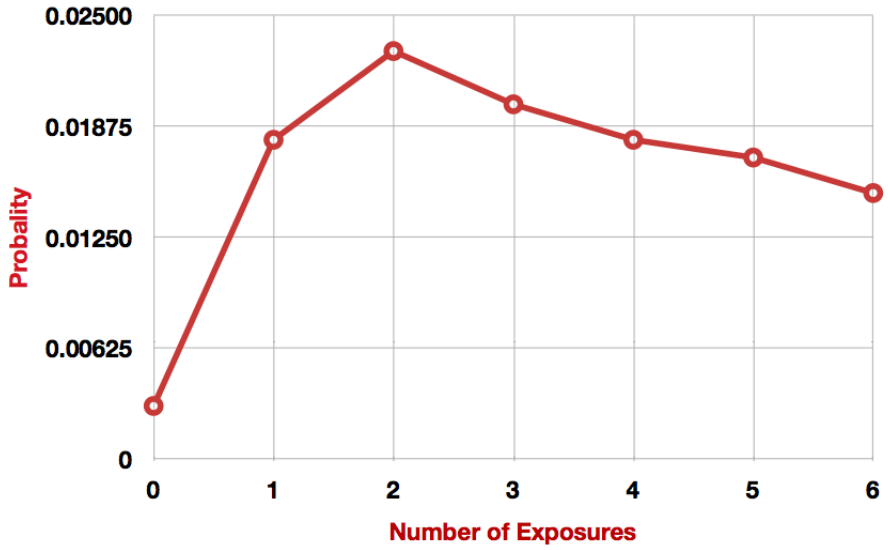

Figure 1: An example exposure curve.

Romero et al found that stickiness determines where the peak of the curve lies and that persistence determines how long the tail of the curve is.

A large amount of other 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.

Romero et al found that different topics are characterized by different exposure curves. [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 [13].

4 Centrality

Centrality is the measurement of how much the network is influenced by a small percentage of nodes. To provide some intuition, Figure 2 and Figure 3 contain example graphs and their centrality measurements. Note that even though there are several different centrality measures, the graphs in Figure 2 and Figure 3 always have the same centrality measurements.

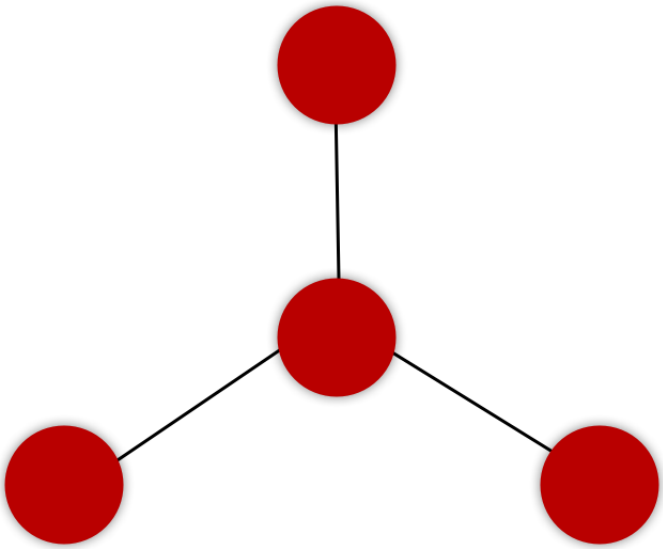

Figure 2: Graph with centrality 1.

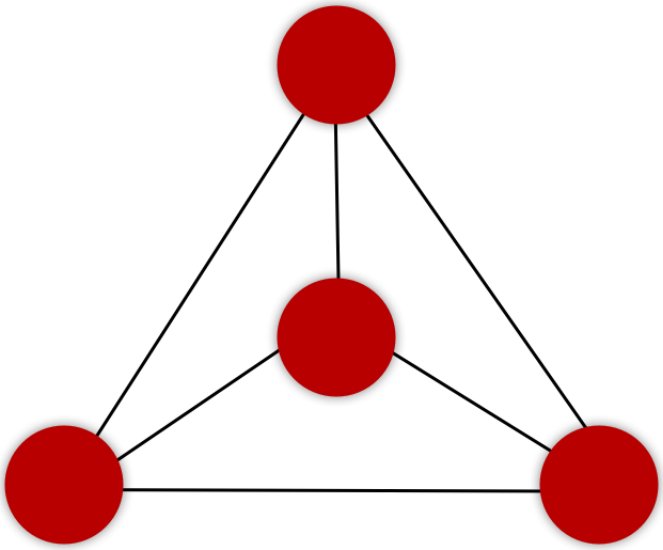

Figure 3: Graph with centrality 0.

For this project, I considered the following centrality measurements:

- degree centrality
- betweenness centrality

- closeness centrality

and ultimately settled on node closeness centrality.

4.1 Degree Centrality

This type of centrality is based on the degrees of the nodes in the graph. It is computed by summing the difference between each node's degree and the max degree of the graph, then dividing by a normalizing constant. Since it only deals with the degrees of the nodes, it is easy to compute, but it is not a very good measure [14].

4.2 Betweenness Centrality

This type of centrality is based on the number of shortest paths each node lies on [14]. Since computing this centrality involves finding and storing all shortest paths (not just all-pairs shortest paths), it is computationally intensive. This is why I chose to not use this measurement.

4.3 Closeness Centrality

The centrality measure that I used is graph closeness centrality which is based on the shortest distances between nodes. It is formally defined as follows [14]:

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

5 Twitter Analysis

5.1 The Data Set

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.

5.2 Case-Study Hashtags

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.

5.3 Results

Using this data, I would calculate the centrality over time, recalculating approximately every 6 days. Figure 4 shows the graph of centrality over time of three of the hashtags.

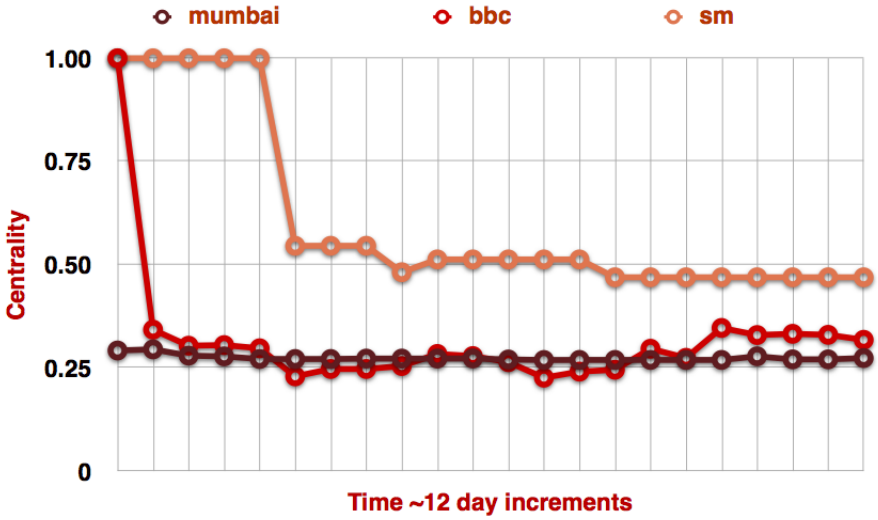

Figure 4: Centrality over time of three hashtags

I have 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.

We can observe some of these problems in Figure 4. Clearly, the curve for mumbai is very different from the other two. Had data been recorded earlier, however, it is conceivable that the curve for mumbai could look similar to the curve for the other two.

6 Theoretical Models

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:

- Small world model
- Erdos-Renyi model
- Barabasi-Albert model

The graph in Figure 5 shows the exposure curves that were used in the simulations. These curves were taken from the results of Romero et al, as discussed in "Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter." The average curve is the average exposure curve over the top 500 Twitter hashtags. The endogenous curve, which is supposed to represent hashtags that deal with ideas that originated in the Twitter network, is the exposure curve for the average of at least 20 hashtags which were classified as being "Idioms." This was a topical category defined in the paper as

> A tag representing a conversational theme on twitter, consisting of a concatenation of at least two common words. The concatenation cant include names of people or places, and the full phrase cant be a proper noun in itself (e.g. a title of a song/movie/organization). Names of days are allowed in the concatenation, because of the the Twitter convention of forming hashtags involving names of days (e.g. MusicMonday). Abbreviations are allowed only if the full form also appears as a top hashtag (so this rules out hashtags including omg, wtf, lol, nsfw).

The exogenous curve, which is supposed to represent ideas that originated outside of the Twitter network, is the average exposure curve of at least 20 hashtags which were classified as being technology-related. This curve was chosen instead of others (such as the curb for the celebrity topic) because it was shaped similar to other topics which were exogenous in nature, but it had more data points.

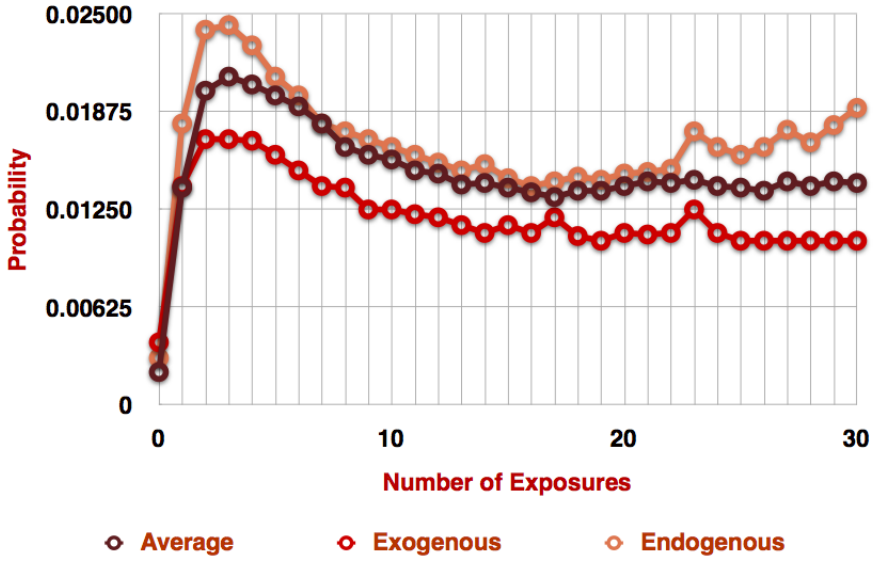

Figure 5: Exposure curves used in simulations.

6.1 Small World Graph

This is 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 $1/d(i, j)^2$.

6.2 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

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

6.3 Barabasi-Albert

This is a power-law graph. We grow the graph by starting with a connected graph (we will start with a grid), 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.

6.4 Results

For each model type, I created a random graph with 10,000 nodes, and approximately $n \log n$ edges. From each of these random graphs, I created three subgraphs, one for each of the exposure curves.

An iteration is as a single pass through all the nodes in the given graph where each node is added to the subgraph based on the relevant exposure curve, where the number of exposures is the number of neighbors of the node which have already been added to the subgraph.

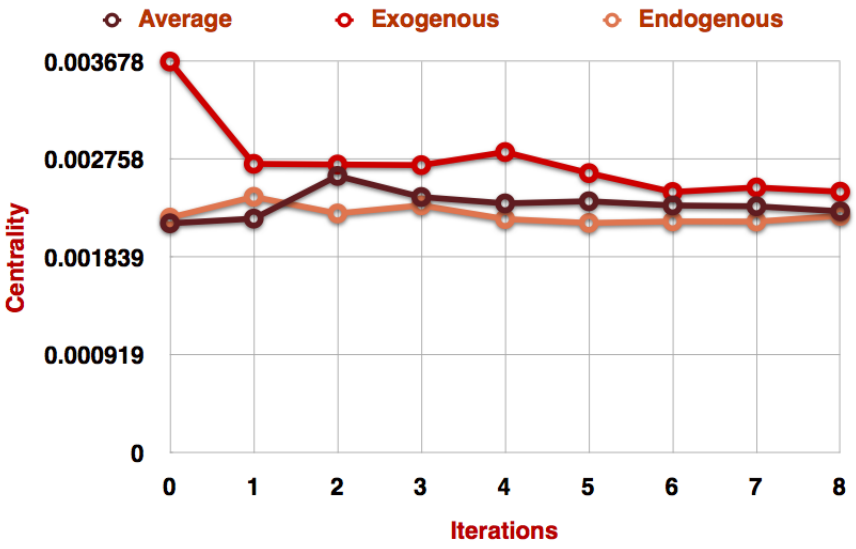

Figure 6: Results of the small world graph simulations.

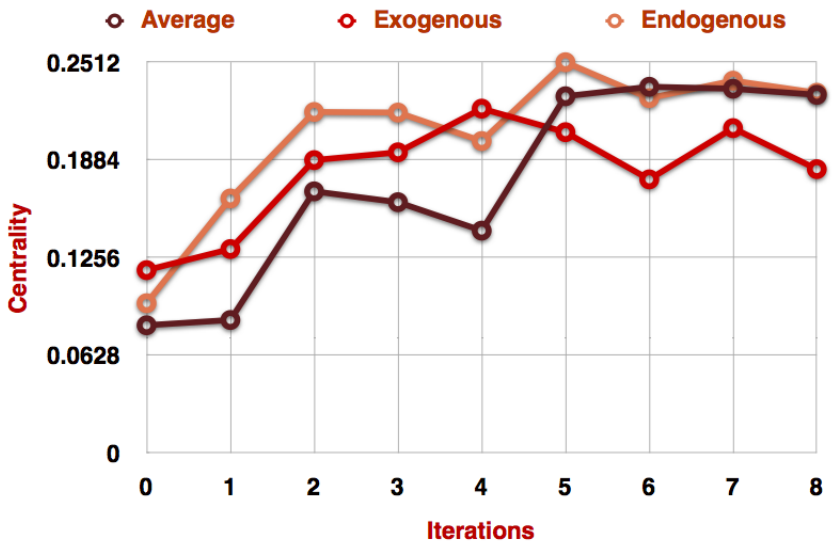

Figure 7: Results of the Barabasi-Albert graph simulations.

Figure 6 shows that in the small world graph, the exogenous information was in general more central than the endogenous information.

Experiments on the Erdos-Renyi graph resulted in all subgraphs having centrality 0, even after 10,000 time steps. This is probably due to the fact that Erdos-Reyni, unlike the other two models, does not begin with a connected graph, and so the random graph from which the subgraphs grew was likely disconnected. This results in very small k values for each node, and very disconnected subgraphs.

Figure 7 shows that in the Barabasi-Albert model, the endogenous information tends to be more centralized than the exogenous information.

Based on the shape of the curves, it appears that the small world graph is more similar to the actual Twitter data than the Barabasi-Albert or Erdos-Renyi models.

There does not appear to be a canonical centrality curve for the spread of hashtags.

7 Future Work

7.1 Centrality Measurement

It would be interesting what the results would look like if we had used a different centrality measurement. If we used degree centrality, would the results be vastly different? I did not consider using Eigenvector centrality initially because it involves find eigenvalues. The best algorithm for this can take a long time to converge. However, it could be the case that social graphs tend to converge quickly, and that this measurement would be both fast and accurate.

7.2 Component Analysis

"Patterns on the Connected Components of Terabyte-Scale Graphs" identifies all of the components as a 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 [13]. One point of future work is to see how this effects centrality– when two components merge, how does the centrality change? Also, is there a correlation between rebel probability and the centrality of the component?

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.

[14] Science of the Web. Luis von Ahn and Brendan Meeder. August 2011. Carnegie Mellon University. December 2011. <scienceoftheweb.org>.