The Memespread Project:

An Initial Analysis of the Contagious Nature of Information in Online Networks

Samuel Arbesman

April 28th, 2004

1. Introduction and Scientific Background

Network structure research is currently a large and vibrant field. Various structural similarities have been discovered among many different types of networks. Some examples of these networks are actor relationships, scientific collaborations, power grids, and the World Wide Web. All of these networks exhibit similarities including the so-called "six degrees of separation," which is due to the structural property of being *scale-free*. When a network is scale-free, it exhibits the property that a few nodes are extremely highly linked to other nodes, while the vast majority of the nodes are not so well-connected. This is similar to what is evident in social networks and is a reason for the apparent small-world nature of the network (i.e. even though the network is large, the distance between any two nodes is small).

Coupled with the study of network structure is the study of flow through these networks. This could include the analysis of the spread of diseases as well as the spread of information. The latter, that of information, is extremely interesting, especially due to the advent of the Internet.

Information that spreads from person to person has been characterized in many ways. One of the most well-known ways was by Richard Dawkins, who in *The Selfish Gene*, coined the term *meme*. A meme was meant to be the cultural equivalent of a gene: both are units of transmission and evolution. While some use it in a narrower sense of only some type of thought-process or concept, “meme” can also be used to include any piece of information that may be passed from person to person.

So what would be memes, broadly viewed? Examples of memes include popular tunes of songs that get stuck in your head, urban legends, jokes, religious beliefs, or favorite aphorisms. Some online memes might be popular links that are passed around, Flash animations that people like, or software or websites that are reached for first when the decision is made to complete a specific task (e.g., if you want to purchase a book online, Amazon.com pops into your head). In sum, any unit of information that is passed on may be viewed as a meme, or contagious information. Of course, when looked at this way, the analogy to a gene is clearly a weak one, since genes have specifically defined biological content (a defined portion of DNA with a specific sequence of bases), and information cannot be similarly quantified. For example, what is the distinct unit of information that is being replicated if the meme is a popular tune: Ten notes? Seven notes? The choice is unclear and memes should not be compared too closely with genes.

Much study has gone into looking at what sorts of memes propagate best, and how they spread throughout networks. Two books on this topic are *The Tipping Point* by Malcolm Gladwell and *Unleashing the Ideavirus* by Seth Godin. They analyze how to best construct a meme so that it is “sticky” (has the best chance of staying in one’s mind and being propagated), as well as how to best insert the meme into the network.

The online equivalent of such a network where ideas can be spread is the community of weblogs that has sprouted up in the past few years, also known as blogspace or the blogosphere. A study was conducted recently by the HP Information Dynamics Lab in which the researchers examined a corpus of compiled historical data (via a web crawler) from a large number of blogs over a period of a month. Using these data, networks of relatedness were inferred based on when certain URLs and descriptions of websites (proxies for memes) appeared on various blogs.

While a retrospective analysis is beneficial, what would be intriguing is an experimental study where a meme is injected into the blogosphere (preferably at a single location) and recorded as it spreads throughout the network of blogs. Such was the goal of the Memespread Project. The Memespread Project aimed to inject a meme, in this case the project itself! The Project consisted of a single website that described the aims of the Memespread Project, as well as an encouragement of the viewers to spread the link as widely as possible.

2. Materials and Methods

The Memespread Project site (found at http://www.arbesman.net/meme.php) was initially linked to a single weblog, kottke.org. Kottke.org, by Jason Kottke, is a popular blog and is listed as number 25 on the Top 100 Popular Blogs, according to technorati.com (as of April 20, 2004). Kottke.org might be what is termed in epidemiology a superspreader, a node that is highly contagious, in that memes that it "incubates" spread extremely widely (Jason Kottke is an informational equivalent of Typhoid Mary).

The technology of the website is relatively simple. The page is a PHP script, that, when it is viewed, logs the date and time, IP address of the viewer, and referring URL (if any) to a separate file. The dates and times that were collected began April 7, 2004 at 22:45 GMT and ended April 20, 2004 at 19:03 GMT. For analysis, the cutoff date and time used was April 14, 2004 23:54 GMT. The data were run through various Python scripts to create graphable data, which was plotted using Gnuplot on an iMac.

3. Results

At its peak, the Memespread Project was the number three piece of contagious information listed on Blogdex (http://blogdex.net). In addition, "memespread" was the second most popular *word burst* listed on Daypop (http://www.daypop.com). The statistical log data were analyzed and were graphed as a histogram of hit count versus time with various bin sizes. The bin sizes used were 1 hour, 1 minute, and 1 day. The histogram graphs are seen below:

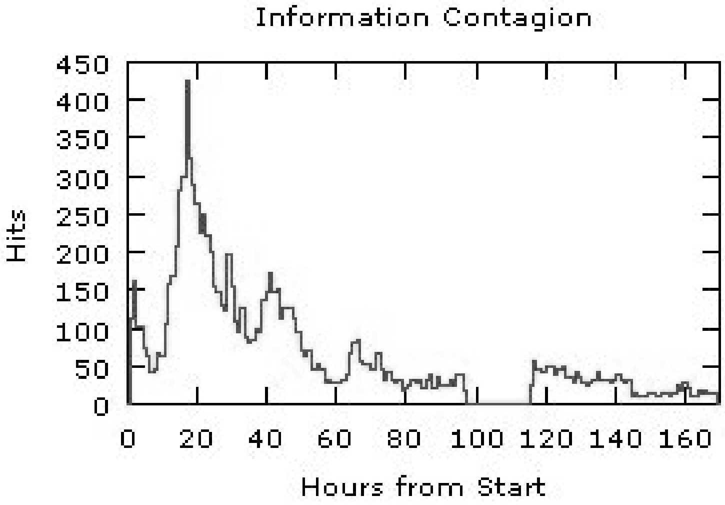

Figure 1. Histogram of hit counts versus time with bin size of 1 hour. While the graph lists zero as the first bin, the first actual bin calculated for is 1 (hence the initial rise from zero, although this is more visible on the day graph). The hit values of zero near the middle of the graph are due to the crash of the arbesman.net server.

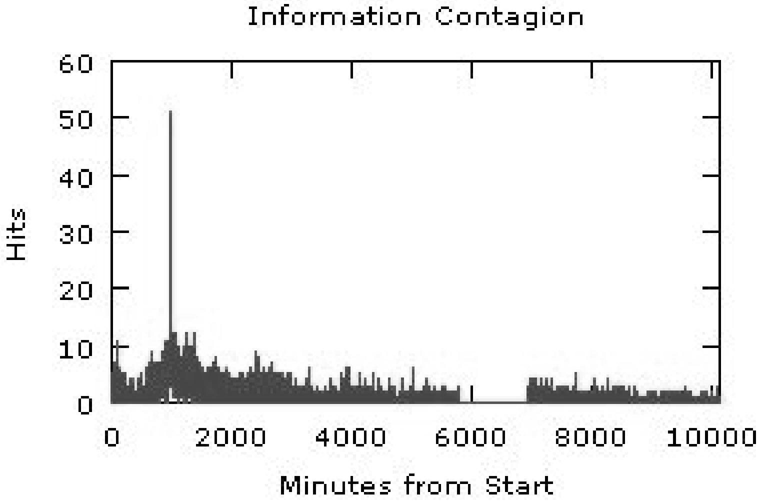

Figure 2. Histogram of hit counts versus time with bin size of 1 minute.

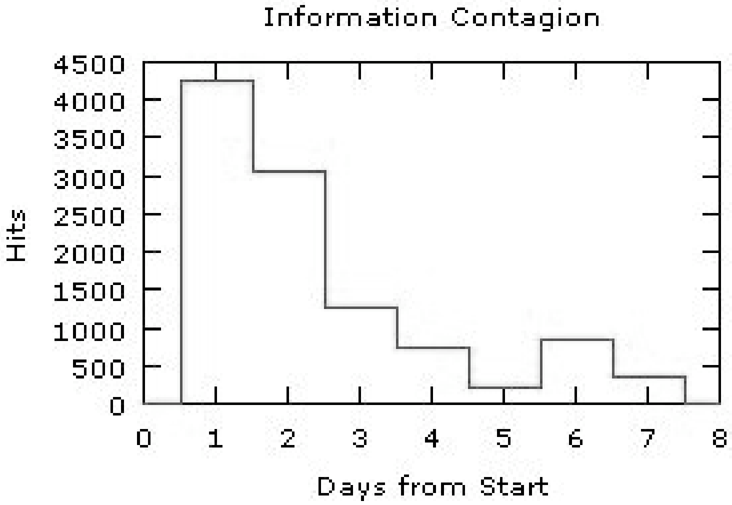

Figure 3. Histogram of hit counts versus time with bin size of 1 day.

Next, a graph was created which showed the cumulative number of hits versus hours from the initial hit. This graph can be seen below:

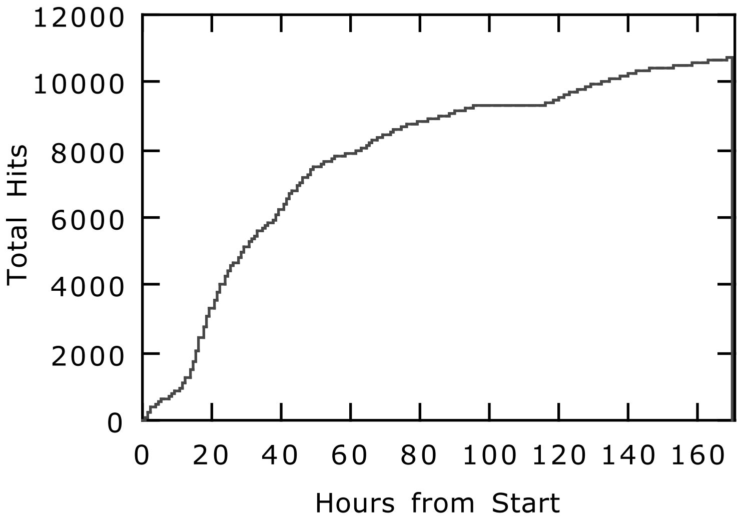

Figure 4. Cumulative hits versus hour from start.

Then, an analysis of the appearance of domain names was performed. A value for each unique second-level domain name was created (e.g. http://www.kottke.org/), whereby each domain name was timestamped with the time of its initial appearance in the dataset and given a scalar value that corresponded to the total number of times the domain name referred a viewer to the Memespread Project. These values were then plotted as spikes on a time axis, overlaid on the histogram of hits versus hour. The graph of these data can be seen below:

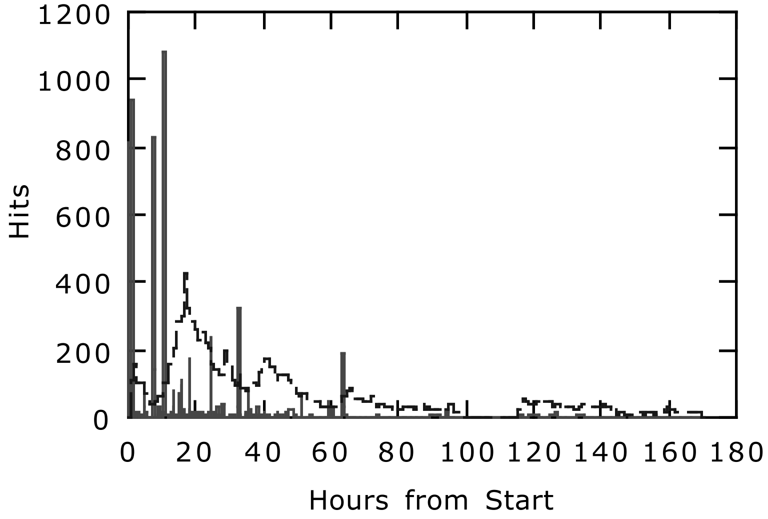

Figure 5. Dark grey spikes correspond to values of second-level domain names and total number of hits on a time axis (spike location refers to initial appearance of the domain name in the dataset). Dashed black line corresponds to the histogram of hits versus hour.

4. Analysis and Conclusions

The initial spike in hits to the website of the Memespread Project, along with a slower decay, fits one of the epidemic profiles discussed by Adar et al. The profile that it seems to most closely fit is that of the "Slashdot effect," so-called because this pattern emerges for those sites that are featured on the collaborative weblog Slashdot (http://slashdot.org). What seemed most interesting is that the largest spike occurred 10 hours into the experiment, and was due to the website MetaFilter (http://www.metafilter.com). The spread of the meme to this collaborative blog seemed to help give the epidemic another wave of spreading (as can be seen in Figure 5 and Figure 1). Of course, some of the peaks and troughs could be due to the cyclical nature of the

Internet, i.e. that there are certain peak usage times which would contribute to these peaks, without any need for an explanation due to the meme being picked up by a blog. The hit peaks also often seemed to follow with a certain delay after the spikes of the initial appearance of the blogs, which is reasonable since not everyone is reading all blogs continuously.

One source of bias in this experiment is what is termed the Hawthorne Effect. This effect is that a study is affected by the subjects' knowledge that they are being observed. Since those who spread the meme of the Memespread Project knew that it was being recorded, they acted differently than if they had acted without the knowledge of being observed. However, it is unclear entirely how the Hawthorne Effect affects the spread of information, but it must at least be acknowledged in any analysis. In addition, one source of error is that, when looking at websites that had the Memespread Project on them, I inadvertently "voted" to raise its appearance on the Plastic.com site (I did not realize that this was what the link I clicked on actually did). This seems to have had little effect however. In addition, while I took careful pains to not tell any of my friends about the experiment and drive them to the site, this is never a perfect endeavor. Lastly, the web server for arbesman.net unexpectedly went down for a period of about 18 hours. This necessarily slowed the spread of the meme. Luckily, the meme was already on its wane, so this was not a great concern overall.

There are clearly further directions to analyze the data collected in this simple study. Eliminating duplicate IP addresses could help gain a better picture of the number of unique viewers to the site. In addition, coupling the IP addresses with their geographical locations (using such a tool as NetGeo found at http://netgeo.caida.org/perl/netgeo.cgi) could be used to create a global map of the reach and spread of the Memespread Project.

Furthermore, the location of the spikes and their height used in the analysis of Figure 5 loses data and might be providing an inaccurate picture of the various referring websites. A way of displaying the importance of each website over time, and not simply data telescoped into a single spike found at its first appearance in the dataset, would provide additional information about how the meme spread through blogspace. Along these lines, counting sites with and without the initial "www" in the addresses as the same would help eliminate errors in the experiment. Along these lines, examining full web addresses might also yield additional results (such as accounting for the fact that LiveJournal URLs are not necessarily referring viewers from the same page). In addition, performing statistical tests on the relationship of hit peaks and the appearance of the websites in the dataset would be insightful.

Furthermore, additional studies might have ways of tracking the informational connections between blogs, rather than through inference, which is all that is possible with the dataset available. This might help confirm the presence of superspreaders, which would have interesting ramifications in epidemiology, where the presence of superspreaders for various disease outbreaks is debated.

Overall, the "informational epidemic" of the Memespread Project was a powerful, although relatively short-lived, example of how information spreads through the blogosphere. It illustrates that data can be gathered about how these epidemics occur and proliferate. In addition, the Memespread Project demonstrates an unusual willingness of individuals to spread a meme, especially if they feel that it is for the sake of scientific research (this was based on reading blogs that mentioned the Memespread Project). Therefore, the Memespread Project is an interesting initial attempt to inject a meme into the blogosphere and analyze its spread. This is clearly a fertile area for further research and should be investigated.

5. Selected Bibliography

Adar, Eytan et al. Implicit Structure and the Dynamics of Blogspace. Available online: http://www.hpl.hp.com/research/idl/papers/blogs/index.html

Gladwell, Malcom. *The Tipping Point: How little things can make a big difference*. 2000.

Godin, Seth. *Unleashing the Ideavirus*. 2001.

Thompson, Nicholas. The Myth of the Superspreader. Available online: http://www.newamerica.net/index.cfm?pg=article&pubID=1214