EVOLUTION OF INFORMATION FLOW ON THE INTERNET: PROJECT MILESTONE

KEVIN LEUNG

As internet usage has grown recently, it has become important in our lives and has affected the way that we act and think. Particularly, people today consume and produce large amounts of media on the internet and in ways very different from before. A common complaint about this change, however, is that we experience an information overload, with status updates constantly coming in from social networks, such as Twitter or Facebook, in small bursts relating to minor events. Some claim that there's a cost to this continuous flow of relatively unimportant information: we no longer pay attention to the big ideas or big movements in the world because we're stuck in the minutia [1] [4]. By looking at the flow of information through social bookmarks on the internet as a graph over time, we can see how valid this claim is.

1. Related Work

Kleinberg [5] describes information streams as having bursty behavior, where individual events grow and disappear quickly . Kumar et al. [6] applied this behavior specifically to the blogosphere. They looked at the evolution of the blogosphere into small communities and characterized its bursty behavior. Their time graph was able to also track the evolution of the network and look at when bursts happen.

Gruhl et al. [3] looked specifically at how information spread through the blogosphere. By focusing on specific topics, they distinguished between the regular, long-running "chatter" around certain topics versus a "spike" from specific events. They also focus in on individuals to see how they behaved.

Leskovec et al. [7] looked at cascades that occurred in the blogosphere. They discovered that most cascades are small, and these tended to be more wide than deep. In contrast to the last paper, however, they did not find bursty behavior across the blogosphere. Additionally, they also propose the *Cascade generation model* to model the data they found.

2. Data Collection

For this project, I decided to use data from delicious, a social bookmarking site. Since I'm interested in the long-term evolution of behavior and long-range effects, it was important to have a dataset spanning several years, with the source of the bookmark as well as the time of the bookmark. Görlitz et al. [2] has a corpus[1] with about 70 million bookmarks

[1] https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/DataSets/PINTSExperimentsDataSets/

spanning from January 2003 through December 2006. The data included the date of the bookmark, the tags associated with the bookmark (which I discarded), an identifier for the bookmark target URL, and an identifier for the user. Unfortunately, because the URL itself wasn't included, I was unable to find any further information about the webpages, such as the type (commercial, blog, etc.) or the original time of creation. Particularly, the first bookmark associated with each node is effectively treated as the creation time of that cascade, though it might be long after the webpage was created.

To compare activity over time efficiently, I split the graph into 4 segments, each corresponding to a year. Between each year, the number of bookmarks increased significantly, jumping by an order of magnitude for the first 3 years:

Year	Bookmark Count
2003	106549
2004	1833989
2005	16120787
2006	54627807

One consequence of splitting these bookmarks by year is that some URLs may be split across years and treated as 2 separate cascades. Since the data doesn't exist before 2003 and isn't available after 2006, this seemed like a consistent, if unfortunate, compromise to make. In the analysis below, cascades of size 1 were removed since they weren't relevant.

3. Results

3.1. **Summary Statistics.**

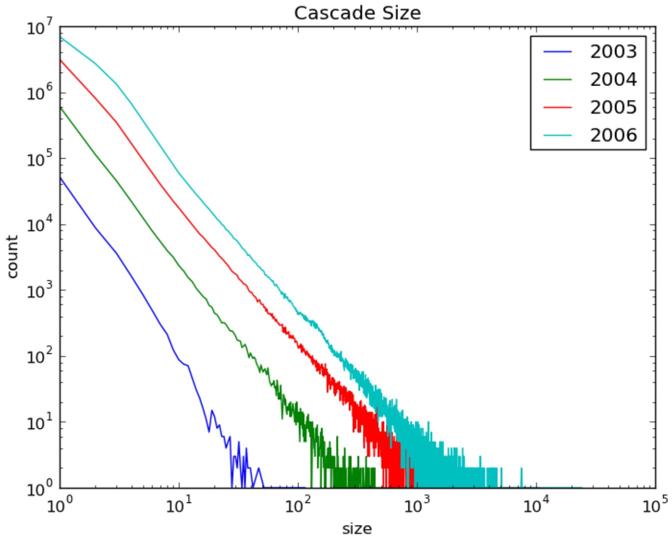
The initial analysis on the data was to see if there was any difference between the years at all. Although the number of networks increased from year to year, it's also possible that the network scaled perfectly with it. First, I looked at the distribution over the bookmark counts for particular bookmark targets:

Plotted on a log-log scale in figure 1, the cascade sizes look roughly linear, suggesting a power law distribution with the αs in the table on the right. These results intuitively make sense: many pages will only be bookmarked once, with larger counts being less likely. Interestingly, this isn't simply a matter of scale: the α also appears to be decreasing from year to year, though at a slower rate. This change suggests that more bookmarks are pointing to the same page. One possible explanation for this is the delicious users have become more social over time and are tending to bookmark more things based on what other are of interest instead of finding pages more randomly.

3.2. **Time Course of Bookmarks.**

Past the size of the cascades, I'm also in the time course of bookmarks and how long these cascades stretch out. To do this, I took the average time difference from the beginning of the cascade to the subsequent nodes. These were both averaged over all cascades and split by cascade size.

To compare how interesting these results were, I also generated a random network. Inspired by random rewiring on times[6], I randomized the times of all of the bookmarks, which should evenly distribute the time overs all of the nodes and kill the bursty behavior.



(A)

Year	Regression α	MLE α
2003	2.75	5.00
2004	2.00	4.03
2005	1.71	3.26
2006	1.66	2.70

(B)

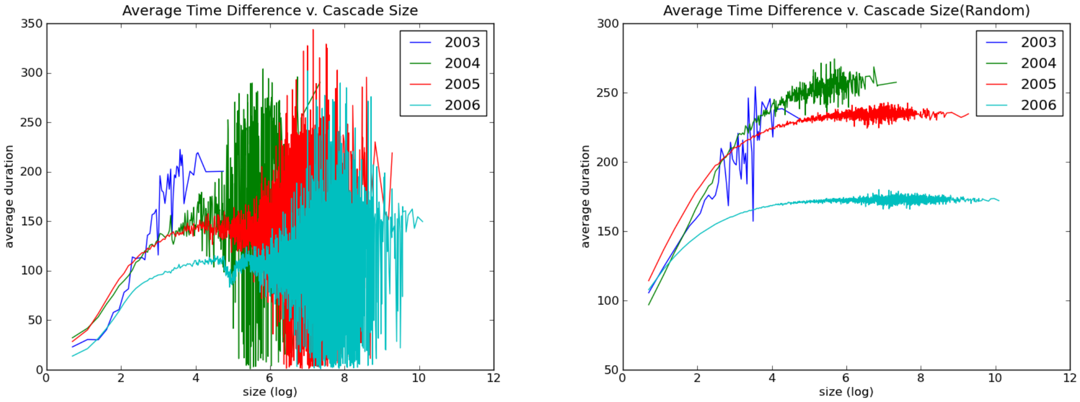
FIGURE 1

This method is fairly crude, and I want to use a slightly better one that retains the starting time for the nodes. Both of these results are below in figure 2.

Interestingly, There's a drop in the average time difference between years in both the real and random data. I immediately don't have an explanation for this result, though my intuition is that it could be related to the year-end truncation of data and the process of ”filling in” the details as the network grows. Looking at the raw data, I noticed very high variance in the average time difference for various cascades of the same size. In the case of a cascade of 2, having one bookmark in January and one in December has a huge effect in pulling up the average.

A good result that we would expect is that the network average time difference is significantly lower than that of the random network average time difference. This confirms our belief that most activity tends to happen early and around the same time instead of being distributed across the year. Again, I'm certain that the average time difference of the random network lying right around 1/3 of the year is no coincidence, but I haven't done the math for it. Unfortunately, there appears to be no other pattern with this data.

In continuing work, I want to compare the results to a generative model of cascades, such as the *Cascade generation model*[7]. Presumably, there's some distribution over the likelihood of different shapes of cascades, and changes in that distribution from year to



(a)

(b)

Year	Network Average	Random Average
2003	32.3	120.87
2004	50.3	122.66
2005	50.79	139.58
2006	30.47	123.58

(c)

Figure 2. Sorry these are line plots instead of scatterplots. I was having difficulty with matplotlib displaying colors

year may not be statistically significant. This model gives us a baseline for how frequently to expect uncommon events and again how that changes with the size of the graph over time.

4. Conclusion

After getting through cleaning the data and looking at the initial statistics, the results are promising in I found some differences to investigate. My primary concern with this particular dataset is that there actually isn't as rich of data as I had originally hoped. Without any understanding of the social structure of delicious or the nature of the webpages, there is little to work with. A major oversight I had with this is that there actually isn't significantly complex network structure: the only connection between various bookmarks is that they're bookmarking the same thing, which makes the cascades somewhat trivial.

It will be somewhat tricky to determine what differences are caused by actual fundamental shifts in behavior versus the effects from scaling a network to a much larger size, though random baseline models should attempt to distinguish that. Given these concerns, however, I believe that I will need to work on more clever analyses to draw out the characteristics I'm looking for in the data. Particularly, my analyses so far have worked around the main issue I'm looking at: how to contrast short, big memes versus long, persisting ideas. Finding the proper metric for that will be critical to the success of this project.

References

[1] N. Gabler. The elusive big idea. *The New York Times*, page SR1, 2011.

[2] O. Görlitz, S. Sizov, and S. Staab. Pints: Peer-to-peer infrastructure for tagging systems. In *Proceedings of the Seventh International Workshop on Peer-to-Peer Systems, IPTPS*, Tampa Bay, USA, 2 2008.

[3] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Infomation diffusion through blogspace. In *SIGKDD Explorations*, volume 6, pages 43–52, 2004.

[4] N. Hassanpour. Media disruption exacerbates revolutionary unrest: Evidence from mubarak's natural experiment. *APSA 2011 Annual Meeting Paper*, 2011.

[5] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2002.

[6] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *WWW '03*, pages 568–576, 2003.

[7] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *Proc. SIAM International Conference on Data Mining*, 2007.