EVOLUTION OF INFORMATION FLOW ON THE INTERNET: PROJECT PROPOSAL

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

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. I am currently considering one of 2 existing corpora. Görlitz et al. [2] has a corpus[1] with about 17 million bookmarks spanning about 4 years. Wetzker et al. [8] has a corpus[2] of 142 million bookmarks collected late 2007 with similar information.

3. Method

3.1. **Corpus Analysis.** The first task will be to sanitize and convert the corpus into a useable form. Wetzker et al. [8] discusses some of the difficulties working with the data, such as bookmark spam. From there, we will need to process it, presumably using a tool such as NetworkX. With a possibly very large dataset, there may be some concerns with how to handle it effectively, though I don't know how to do that yet. The graph itself will be the bookmarks connected by the links in the bookmarks, annotated by the date of creation.

With the dates on the bookmarks, I will look at slices of the network and prefix graphs to see the evolution of the network over time. Since I'm interested in how behavior has changed, the analyses described below will be done on segments of the network, which will then be compiled into statistics over time to be compared.

The bulk of the work will be to track the relationship between the burstiness/locality and the persistence of certain topics. A simple way to think of this is that the minutia cascades of the graph should have short timespans but a wide spread. More inspiring topics should generate cascades with long timespans and wide influence. Obviously, there's a confounding factor here in that both quick memes and grand ideas should generate a large amount of activity (i.e. large cascades).

3.2. **Raw Evaluation.** To analyze this, I want to consider several possible measures. One comes from Gruhl et al. [3], which presents a mathematical model for the difference between chatter and spikes. It's also possible that we can measure these effects using more simple network properties. We might get an intuitive sense of this tradeoff by comparing the average time duration of bookmarks to a webpage against the size of average path length within the cascade generated by the webpage. These statistics can be plotted to see the change over time as well as the frequency of various events. For example, a positive result would be a drift in the ratio of long and short cascades.

3.3. **Comparative Evaluation.** Although the graph of these statistics over time are the primary, interesting result with respect to the original question, I'm also interested in how much this diverges from expected behavior. It's possible that the results are no different than what we may see in a random graph that also experiences

[1] https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/DataSets/PINTSExperimentsDataSets/
[2] http://kmi.tugraz.at/staff/markus/datasets/#[Wetzker 2006]

the same growth as in internet usage. There are two baselines to compare the data to.

First, I'll compare the results to a random graph. One way to do this is to randomly rewire the network[6]. Leaving all of the timestamps on nodes intact, the end points of all edges can randomly be changed to another bookmark which occurred in the past (you can't bookmark a webpage that hasn't been created yet). This model will obviously lose much of the bursty behavior we saw before, but it does give us a sense of how scale alone affects the network.

Second, I'll 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 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

The result of this study will be a 10 page paper that quantitatively describes how the flow of information on the internet has evolved over the course of several years by looking at activity in social bookmarking. Success will be in resolving the question of whether we've seen a rise in short, bursty activity and a decrease in the persistence of big, far-reaching activity. More specifically, the results should quantify how much activity has changed beyond what is expected from growth alone, with possible consequences for how internet behavior has changed.

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.

[8] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *Mining Social Data (MSoDa) Workshop Proceedings*, pages 26–30, 2008.