CS224W REACTION PAPER

KEVIN LEUNG

I'm interested in how the flow of ideas has changed over time as human behavior on the internet has changed. Beyond its proliferation, the internet has changed the way we interact with it and perhaps how we think as well. This topic touches on 2 papers that have discussed the structure of the blogosphere as a graph over time. In this reaction paper, I will summarize the papers, critique them, and brainstorm way to extend ideas or base new ideas from these papers.

1. Kumar et al. 03

1.1. **Summary.** This paper presents the evolution of the blogosphere into small communities (< 20 people) with bursty behavior. For this application, the paper develops the time graph: a graph with durations on its nodes and timestamps on the edges. In this context, the nodes are blogs that exist within a duration, and the edges are links from one post to another that are written at a particular time. From this, they apply a community extraction algorithm, and these communities are the primary unit in further analysis. Within these communities, the authors use the timestamps to find bursts of activity.

In their results, the authors note a few significant changes in the structure of the network over time. Near the end of 2001 through 2002, the largest SCC grows from 3% to 20%, a large change in the global structure of the network. At the same time, the fraction of networks joining communities increases as well, a change int he local structure of the network. Compared to a random network (where all edges are the same except with new random endpoints), the fraction in communities are higher than expected while the growth of the SCC is lower than expected. The authors also note bursty activity in the communities, with a sharp increase in burstiness at the end of 2001 that isn't explained simply by the growth of communities.

1.2. **Critique.** Altogether, the paper puts forth an interesting picture of the evolution of networks, combining a new type of model (the time graph) with several existing analysis techniques and algorithms to come to several observations. I, however, do have 2 critiques of this paper to point out.

First, they remove nodes (blogs) with large in-degrees. They explain that they're trying to characterize communities and should omit this different sort of activity, but this also gives a somewhat skewed perspective on the structure of the blogosphere. Although they may have captured the behavior of small communities, there presumably is more interesting activity on a larger scale that they have chosen to ignore.

Second, they claim that the burstiness of the communities had increased in a way that wasn't explained by the growth of communities based on an increase in average burstiness per community. This claim, however, treats the communities in isolation. This increased activity may have been based on weaker connections between different communities, the number of which have been growing. Similar to my first critique, this critique is largely aimed at how the authors don't reconcile their findings on communities with the macro state of the network.

2. Leskovec et al. 07

2.1. **Summary.** This paper investigates cascades within the blogosphere over time, trying to characterize the frequency of various types of cascades and creating a generating model of it. Instead of working with a time graph like the last paper, this paper splits blogs and posts into distinct graphs for different analyses. The blog graph has weighted edges between blogs for the number of links, and the post graph uses posts as nodes and adds directed edges with δs between the post times. In relation to the first paper and the premise of this reaction, I'll focus primarily on the post graph, which can be broken down into cascades.

The first result is that posts and links are periodic over the week. After fitting the data to a power-law distributions, the authors found that the observed bursty behavior is simply a product of optimal timing. This fit of a post's popularity to a power-law distribution is particularly significant. Another notable result is that only 1.2% of the cascades are more interesting than isolated or single edges, and of the remaining, most tend to be more wide than deep. To explain these results, the paper proposes a *Cascade generation model*, which captures many of the characteristics of cascades in the data with only a single parameter.

2.2. **Critique.** Overall, this paper offers more general observations about the structure of the blogosphere, particularly with respect to the relationship of various characteristics to the power-law. It also gives a good characterization of cascades within the network, though it seems to disregard what I consider the interesting outliers. It

claims that long chains are rare and also points out outliers from the predicted relationship given. I think these oddities deserve more attention. Although the data is a good fit to the model disregarding these, the unique characteristics of the blogosphere should emerge more noticeably from differences from standard models.

In relation to the first paper, the dismissal of bursty behavior is odd. Within the results, it only goes as far as to attribute the bursty behavior to another source, though doesn't appear to directly deny its existence. The introduction, however, makes a stronger claim that "we found that blog posts do *not* have a bursty behavior." These details should be reconciled as the purpose of these claims within the scope of the paper isn't clear.

3. Synthesis and Brainstorming

A major inconsistency I noticed between these 2 papers is that the first paper (Kumar et al. 03) makes a strong claim about the burstiness of the blogosphere, where the second paper (Leskovec et al. 07) claims that there is no bursty behavior and is instead primarily periodic behavior. One reason why this may have occurred is that the two papers were looking at network behavior on different scales. The first was primarily focused on communities, and the second took a larger perspective. It's possible, then, that bursty behavior is only found in communities.

Another might be a difference in the corpora. The first paper took data from 1999 to 2002, and they noticed tremendous structural differences within months. The second paper took data from 2005, and given how quickly internet culture has changed, it's possible that individual behavior may have changed enough that bursty behavior could have disappeared.

This issue could be pursued by a longitudinal study of network evolution in the blogosphere over several years, looking for more fundamental changes that aren't apparent from this disparate slices of time. We can apply the some combination of the analyses above on this new data set.

With a longer period of time, it could also be more worthwhile to look for larger cascades and any regularities in that. Leskovec et al. (07) claimed that "cascading behavior appeared relatively rare," though they do attribute this to possible data issues. Given that, it could be helpful to compare short cascades and bursty behavior with longer cascades.

To broaden this idea, a common concern with the proliferation of the internet is "information overload," where humans are unable to develop or maintain lasting ideas because we're constantly being presented with short-lived details or other minutia.

Within the framework of these papers, this concern predicts that over the past 10 years or so, we should observe an increase in short cascades and bursty behavior, with a correlated decrease in the time duration of cascades. This prediction seems very testable and extends the observations of these 2 papers into a more general claim.

4. References

R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW 03*, pages 568576. ACM Press, 2003.

J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, M. Hurst. Cascading Behavior in Large Blog Graphs. In *Proc. SIAM International Conference on Data Mining, 2007.*