1 Introduction

One topic of current interest is how language and sentiments are transmitted in the blogosphere and how individuals affect the sentiments of those in their online sphere of influence. Therefore, the focus of this reaction paper is threefold: (1) to summarize four journal papers related to sentiments and influence, (2) to critique the methods and scope of the journal papers and (3) to brainstorm directions for the term project.

2 Literature review

2.1 Java et al. 2006

Java et al. (2006) explore models that determine the blogs with the most influence on the Blogosphere. The paper defines influence as a link from blog $a$ to blog $b$ implies that $a$ influences blog $b$. They further create an influence graph that weights edges by a function of the amount of links. This paper evaluates various heuristics, such as Pagerank, indegree, and greedy algorithm, to determine which blogs have the most influence on the Blogosphere. The analysis shows that PageRank is rather efficient and converges quickly.

The paper also analyzes the effect of splogs, i.e. spam blogs. The results indicate that not removing splogs greatly impacts the accuracy of the influence models. As such, the paper uses some algorithms to identify splogs, and then, and only then, do the heuristics converge to within 70% agreement on the predicted number of influenced nodes.

2.2 Leskovec et al. 2007

"How do blogs cite and influence each other? How do such links evolve? Does the popularity of old blog posts drop exponentially with time?" These questions are addressed by Leskovec
This study creates a directed graphs to analyze shape patterns in the large blog graphs as well as temporal patterns regarding how the popularity of a blog post decays with time. Specifically, the paper develops two graphs, the first, the "Post network", and the second, the "Blog network". The Post network is based on the time of posting and is created by looking for cascades starting at a node with no outlinks and moving up the cascade. The Blog network is defined by linking blogs to each other if and only if there is a blog post link between the two. The edge has a weighting function related to how often blog posts link to each other.

The authors note that 98% of the blogs are singletons and most blog post cascades are short. Nonetheless the paper has enough data to show some trends in the cascade structure of the Blog network, such as the power law distribution of the cascade sizes. In addition, the authors noted the power law decay of the popularity of a blog post when studying the Post network.

2.3 Leskovec et al. 2010

Leskovec et al. (2010) is about predicting positive and negative links in online social networks. In this case, “positive” means something like friendship or agreement, while “negative” means opposition or antagonism. The key idea in the paper is that we can try to predict a person’s attitude towards another person by looking at their relationships with those around them. The paper describes two sets of features for a machine-learning framework that can be used to predict the polarity of these unknown links, and offers evaluations of how well these sets of features perform. The paper also makes parallels to social psychology’s concepts of balance (“the friend of my friend is my friend” and all the similar variants) and status, both at a local and global scale.

The two feature sets that the paper uses for machine learning are the degrees of the node (the cross of positive/negative and incoming/outgoing) and a triad set, which represents the positions a node has in the triads it is part of (its relationship with both other nodes, and the relationship between those two others, all with pos/neg polarity). It is interesting to note that both sets of features extend to only one node away on the network (and, of course, the connection between those first-degree nodes)–they are relatively local features.

The main technical approach is a logistic regression classifier with either the degree, triad features, or both, compared against a random baseline. Across all datasets (wikipedia, opinions, and slashdot) the results were fairly stable (with wikipedia being a slight outlier, which is hypothesized to be because of its open election environment where decisions (corresponding to pos/neg links) are visible to all nodes), suggesting that the findings are general in nature. These findings were that degree and triad features both significantly improve edge polarity prediction, with the triad features being slightly more powerful, and that both can work well together but for much smaller gains. The paper further tests the effects of including information about negative edges versus excluding that information, and finds that the inclusion of negative information offers a 50% increased boost (on top of positive) over random.

The paper finds that balance properties are relatively generic across the different datasets,
while status depends more on the particular dataset. In terms of scale, both results hold fairly well locally. Globally, if balance were to hold, the prediction would be something akin to two mutually opposed factions (or one large group of friends), and if status were to hold, there would be a global ordering. The paper finds little evidence for a global balance structure, but does find evidence for a global status ordering. (This is relative, rather than strict: there is much more ordering than a random network with the same properties otherwise.)

2.4 Java 2007

This paper is essentially a short framing of the problem of modeling influence, opinions, and structure of social networks. Java proposes to model influence using topics, polarity of sentiment, bias, and temporality. Java also notes that there are several challenging questions that this type of work hopes to address: chiefly, that blogs and other social media contain noisy, ungrammatical, and poorly structured text, and that contractions, negations, and sarcasm make shallow NLP ineffective and demand instead a complex, semantic understanding of the text.

One of the interesting arguments made by Java in this paper is that sentiment and opinions are important components for understanding influence. Java claims to introduce the idea of ‘link polarity’, which is somewhat different from the link polarity discussed in Leskovec 2010. In particular, the type of influence Java mentions is that if A links to B with negative sentiment, then (he argues) influencing B would have little effect on A. In this sense, link polarity is used not necessarily as an open-ended ordering (akin to the status concept discussed in Leskovec 2010, which might be applied to many interactions), but specifically as an influence hierarchy. Java also mentions that there is likely a ‘tipping point’ at which an aggregation of opinions over many users begins to influence new users.

3 Discussion

Each paper focuses on different aspects of network propagation and structure, sometimes improving one part of the model while providing simplifying assumptions on other model aspects.

Leskovec et al. (2007) note in their introduction that inferring relationships between posts from text is inaccurate. Therefore this paper considers only direct propagation of links. However, it is possible that the blog network edges should be weighted not only by the number of back-and-forth blog post links, but also by clues from the text. As Java et al. (2006) note, emotions, for example, might affect the influence of a blog post. As text-tagging tools become more sophisticated, including textual clues might more accurately describe the edge weights in the graph of the blog network in both Java et al. (2006) and Leskovec et al. (2007).

Another point of difference is how each paper cleans the data. While Leskovec et al. (2007) were careful to include only popular blogs, Java et al. (2006) included all blogs
except for splogs, which they removed. Their paper shows the sensitivity of the results to the inclusion of splogs and mention algorithms to remove these blogs from the data set.

One perspective missing from the different papers discussed is an event-based study. Each shows how influence or information flows in general. However, what might be the specific characteristics of flow after a given type of event, such as posting of a foreign policy news item or a change of relationship status?

Java et al. (2006) pose the question about how the influence model should be modified to include a blog's existing emotions. The authors mention the example of a fictitious iloveipods.com site, which would have little impact on its followers to buy iPods, since presumably, most followers would already love iPods and own at least one. The idea of weighting edges based on the existing emotions on a topic, as mentioned in the paper, could be extended to the more general idea of mood variance. In other words, does a sudden change of the average mood of a blog cause that particular blog post to have a greater influence?

One potential area of improvement or further research in the direction of Leskovec et al. (2010) would be to look at what precisely positive and negative links indicate in terms of social meaning. They work with Epinions, Slashdot, and Wikipedia data, where positive and negative are trust/distrust, friend/foe, and supporting/opposing vote (for moderator), respectively. While these do clearly fall into the categories of positive and negative, there is a distinction between declaring someone a “foe” and opposing a promotion for which they are being considered. In the same vein, one might consider looking at a finer granularity of signed links, specifically someone “liking” a post on Facebook. The issue raised there is that many links in social networks tend to be only positive, or at least only positive at a superficial level. With some deeper understanding, it might be possible to extract more precise or nuanced polarity information from a link. For example, if we could extract sentiment information (which could go beyond positive/negative and might include supportive, consoling, dismissive, sympathetic) from comments on posts on social networks, we could create signed links where the signs represent a plethora of different information.

A heartening aspect of Leskovec et al. (2010) is that very interesting findings are only one node away on a network. Rather than having to study the local network at, say, 5 or 10 links deep to learn enough to accurately predict link polarity, all that is needed is a complete view of the nodes one link away and their interconnections. Studying structure in this way is appealing because it has a high signal/noise ratio, and it is reassuring to see that non-trivial results are obtainable while looking at the network on a very local scale.

4 Project proposal

The papers we have reviewed suggest at least two possible directions for the term project, as our team brainstormed with Prof. Leskovec and was alluded to in the discussion above. The main concern is to find a subset of data such that the signal is stronger than the noise. Both directions share in common the idea of a static network, within which we hope to track a dynamic process, namely the flow of sentiment.

One issue that we will have to address (which is very data-dependent) is how to model
the network. One possibility would be to consider (if we were working with Facebook-like data) each post as a node, and comments on that post as links between two supernodes (supernodes being users). The polarity of the links in that case could be the sentiment extracted via some challenging NLP. Another way to model the network might have an article (in the blogosphere) or original tweet (in the Twitter world) as a node, and then hyperlinks or retweets would be the nodes, with the text included/surrounding the hyperlink/retweet as representative of the sign of the link. In either case, part of the interesting challenge will be to address the NLP issues of extracting useful sign information from the context.

Another issue is what type of sign information we would like to focus on. Many social networks are overwhelmingly positive (no ‘dislike’ on Facebook, for example), and getting an interesting mix of social meaning may not be possible on a positive/negative sentiment axis. There are other axes we are considering, which include democratic/republican, liberal/conservative, hippie/governmental, religious/atheist, supportive/unsupportive... In short, there are many dimensions along which one might split textual information. We are currently exploring data possibilities, searching for a dataset that will give us an interesting and varied representation. Adamis and Glanc (2005), for example, provide an excellent framework for structuring this type of problem, following politically-oriented memes in the blogosphere. However, they do not also include the surrounding blog text, which we see as an opportunity to use to track sentiment.

In short, we propose to examine the propagation of social meaning through social networks. The particular form of social meaning that we study, as well as the way in which we model our network, will be highly dependent on the dataset(s) we choose to work with. We have highlighted some examples above of possible modeling techniques, as well as suggested some possible dimensions of social meaning that we are interested in studying.

References


