Sentiment Flow Through Hyperlink Networks

Mahalia Miller[†*] **Conal Sathi**[‡*] **Daniel Wiesenthal**[‡*] **Jure Leskovec**[‡] **Christopher Potts**[†]
Stanford University, Stanford, CA 94305 USA
[†]{mahalia, cgpotts}@stanford.edu
[‡]{conal, dw, jure}@cs.stanford.edu

Abstract

How does sentiment flow through hyperlink networks? Earlier work on hyperlink networks has focused on the structure of the network, often modeling posts as nodes in a directed graph in which edges represent hyperlinks. At the same time, sentiment analysis has largely focused on classifying texts in isolation. Here we analyze a large hyperlinked network of mass media and weblog posts to determine how sentiment features of a post affect the sentiment of connected posts and the structure of the network itself. We explore the phenomena of sentiment flow through experiments on a graph containing nearly 8 million nodes and 15 million edges. Our analysis indicates that (1) nodes are strongly influenced by their immediate neighbors, (2) deep cascades lead complex but predictable lives, (3) shallow cascades tend to be objective, and (4) sentiment becomes more polarized as depth increases.

Introduction

The ever-growing amount of data available on the web, predominantly as prose in hyperlinked blogs and social network posts, has inspired much research focused on understanding interactions between authors and the trends emergent in the language they use. Blogs and news articles on the Internet offer an opportunity for readers to rapidly share their thoughts and opinions on an issue by creating a new post that hyperlinks to the original post. In this way, authors create a directed graph in which posts represent nodes and hyperlinks represent edges. By looking at how emotional language differs from post to post, we investigate the ways in which one author influences another's written sentiment.

Previous work on both sentiment and network analysis inspires several areas for exploration in the overlap of these two mostly disparate fields. Does the sentiment of one post influence the sentiment of its immediate neighbors? How does sentiment flow through the blogosphere on a macro scale? Are there noticeable differences in the sentiment characteristics of posts in deep cascades versus shallow cascades? In this paper, we propose a preliminary approach to answering these questions, apply our approach to a large dataset, and present significant results showing clear trends with real-world implications.

[*]The first three authors contributed equally to this work.

Prior Work

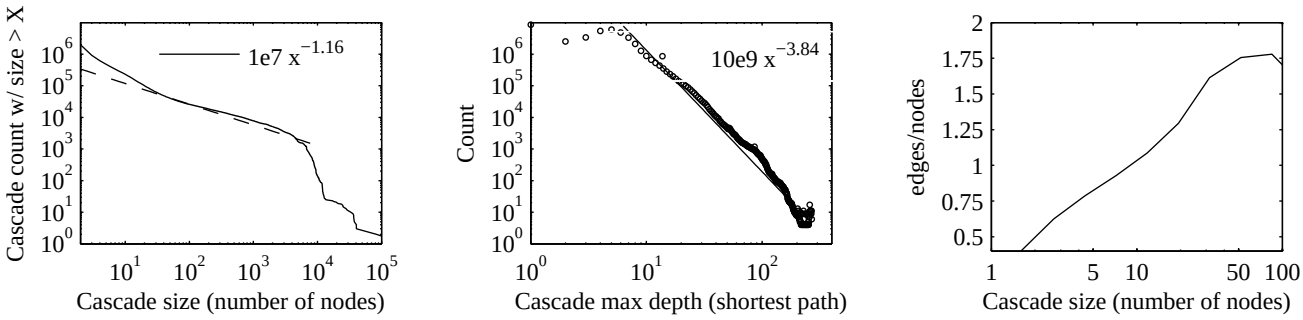
Prior work on blogosphere graphs has explored linking trends and patterns in cascades. Adamic and Glance (2005) examine linking trends and the connectedness of Democratic and Republican blogs. Leskovec et al. (2007) explore patterns in the topological properties of cascades across the blogosphere. Prior work in the field of sentiment analysis attempts to determine the sentiment of individual texts in isolation (Pang, Lee, and Vaithyanathan 2002), or to track the propagation of explicitly labeled sentiment within the context of one social network (Zafarani, Cole, and Liu 2010). This paper combines ideas from the graph analysis and sentiment analysis fields in order to analyze the flow of sentiment in blog post networks connected by hyperlinks. In contrast to previous work that tracks links, phrases, or memes only, our approach is to analyze the full text of a post using rough sentiment heuristics, and to track the flow of these extracted metrics over the linked network.

Methodology

Dataset Preparation

We obtained data from the ongoing MemeTracker project (Leskovec, Backstrom, and Kleinberg 2009) for the month of August 2010, which consists of roughly 1 million blog posts per day. MemeTracker builds maps of the daily news cycle by collecting roughly 900,000 news stories and blog posts per day from 1 million online sources, ranging from mass media to personal blogs. Each post contains a URL, time stamp, and all of the URLs of the posts it cites. We worked with a newer version of the data than that referenced in Leskovec et al. (2009); this new version includes the full text of the webpage (in addition to key quotes), upon which we ran our sentiment analysis.

Our dataset was initially comprised of nearly 1.5TB of raw text representing a full month of MemeTracker data. We pruned out singleton nodes (nodes with 0 in- and out-degree) from our dataset. This allowed us to focus on the interactions we were interested in: non-trivial cascades of at least two nodes in which one node "sees" another and whose sentiment is potentially affected by it. We also removed self edges to the same blog post and edges pointing out of the dataset since they do not show the flow of sentiment. This process of data preparation yielded a large, linguistically rich, highly connected graph of roughly 8 million



(a) Cascade size CCDF (b) Cascade depth distribution (c) Edges vs. nodes

Figure 1: Our dataset is rich for cascades with a size smaller than 10,000 nodes and max depth less than 100.

blog posts and 15 million hyperlinked edges from 4 million blog domains.

Sentiment Extraction

We determined the sentiment of posts in our dataset by combining sentiment scores of all words in the post. We obtained these scores from two established sentiment lexica: the Harvard Inquirer (Stone et al. 1966) and SentiWordNet 3 (Baccianella, Esuli, and Sebastiani 2010). These lexica offer binary and real-valued scores, respectively, for the positivity and negativity of a word; SentiWordNet additionally offers an objectivity score. We used an augmented bag-of-words model similar to that used by Pang et al. (2002), in which we reversed the polarity of words between a negation ({not, *n't, no, never}) and the next end-of-sentence punctuation. The sentiment attributes of a post that we examined–positivity, negativity, and objectivity–were the averages of the scores of the words in the post. Thus, the poem beginning "it's hot as her fervent eyes/ welcoming as her cute smile/ soothing as her sweet words/ calming as her delicate hands" received high positivity and low objectivity scores.[1] SentiWordNet gives good coverage over our data, offering sentiment tags for 35% of our words, and at least one word in 92% of posts. We validated our process internally by ensuring that our findings held using either lexicon. Similarly, while the two are not completely independent–the Inquirer was used to validate seed sets for SentiWordNet–the agreement we found between lexica suggests that our findings are externally robust and not an artifact of a particular lexicon. Our analysis focuses on SentiWordNet since it has higher recall than the Harvard Inquirer and is more nuanced.

We also propose sentiment extraction using emoticon presence. Emoticon presence is binary (a token either is or is not an emoticon), so we assign post-level sentiment scores based on emoticon frequency. Positive emoticons are considered to be [:), :D, :P, :p, ;)] and negative emoticons are [:(, D:]. We do not invert the polarity of an emoticon in relation to negation markers (as we do for words), because emoticons play no role in the compositional semantics of a post and are not scoped by negation. This approach correctly yields, for example, a high emoticon-based negativity score to the post "Boys! Stop! Leave me alone! :(I want one guy but that can't happen right now :(" [2]

To gain a meaningful sense of post sentiment, we first determined the author's usual sentiment scores, and then considered how far the post in question deviates from this baseline. This allowed us to take into account the fact that authors exhibit unique writing styles, and that there is no universal "positive": a crotchety old blogger's less-negative-than-usual post may in fact be his own peculiar way of showing positive sentiment. For example, an often self-deprecating site, mylifeisaverage.com, had higher negativity values than the respected news site nytimes.com (negativity values of 8.47e-2 and 3.34e-2 respectively). Additionally, comparing relative-to-baseline sentiment values allowed us to make valid comparisons between changes in features regardless of lexicon. We made the simplifying assumption that each web domain (e.g. jdoe.blogspot.com, nytimes.com) represents an "author," and averaged across all posts from this domain to obtain baseline values. While nytimes.com may have several contributing authors, for example, it is reasonable to consider a news outlet as presenting a coherent, generalized "voice" or style of writing. Our subsequent analysis is based on deviations from these baselines.

Cascade Identification

We constructed a network graph of our data using the C++ SNAP network analysis library. We modeled the data in the graph as follows: each node represents a blog post with its sentiment scores, and directed edges between nodes represent hyperlinks. An edge points from node u to node v where post u contains a hyperlink that cites post v. We were specifically interested in information propagation graphs (cascades). After constructing the graph, we searched for "cascade initiators" by iterating through each node and looking for those with a non-zero in-degree and a zero out-degree. These nodes represent posts that begin a chain of links. Once cascade initiators were found, we searched for all the nodes in each cascade by applying a breadth-first algorithm, starting at each cascade initiator and following the in-links. This process identified 1.9 million cascades.

[1] Post URL: http://shayrionline.blogspot.com/2010/07/welcoming-as-her-cute-smile.html

[2] Post URL: http://kiim-c.blogspot.com/2010/07/boys-stop-leave-me-alone-i-want-one-guy.html

Cascade Topological Properties

We first examined the topological properties of our cascades, which we treated as disconnected subgraphs. We were interested in understanding basic properties of our data, namely the cascade size and depth at which our data would become sparse, whether we had tree-like cascades, and as a check whether our cascades exhibit similar properties to those in similar datasets.

The complementary cumulative distribution function of the cascades as shown in Figure 1(a) is roughly distributed along a power law distribution with coefficient of -1.16. This finding is in agreement with research on similar data, such as Leskovec et al. (2007) which found a power law coefficient on the CCDF of -1. Figure 1(b) shows that the dataset is rich for the chosen ranges of interest, specifically cascades of maximum shortest path distance less than 35.

While we might have expected tree-like structures, in fact, many of our cascades are instead DAGs that are not balanced trees. If the ratio of edges to nodes in Figure 1(c) were around 1, then the number of edges in a cascade would be increasing almost linearly with the number of nodes and thus the average degree in the cascade would remain constant as the cascade size grows. In our dataset, however, the average degree increases over time, suggesting that larger cascades become more densely connected.

Analysis

Post Level Analysis

Does sentiment flow in a cascade? Figure 2 shows that indeed sentiment does flow from one post to another. When the parent is more objective than normal, the child post is also more objective. Similarly, when the parent gets riled up, the child uses more subjective language. However, the slope of the parent-child subjectivity relationship is less than 1 (merely 0.09) and crosses the origin, which indicates a moderating effect in the parent-child interaction. For example, if a parent is much more subjective than usual, our data shows that the child also uses more subjective language than usual, but the response is tempered.

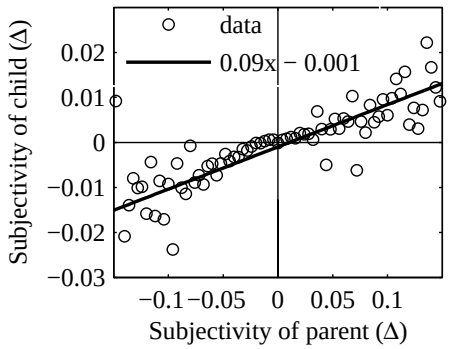


Figure 2: Child subjectivity (Δ) vs. parent subjectivity (Δ).

Cascade Level Analysis

Not only does sentiment flow from one post to another, but our results show that in the context of a whole cascade, sentiment flows with distinctive trends. Sentiment in a cascade exhibits four phases: (1) at the cascade initiator, language is close to baseline, but then (2) positivity and negativity heat up very quickly, (3) cool off just as quickly (though not as much), and finally (4) return slowly to a lukewarm baseline. Figure 3 shows these clear trends of sentiment values (relative to baseline) as a function of a node's distance from its cascade initiator. The cascade initiator's objectivity ($x = 0$) is close to baseline (1). Sentiments quickly become heated, with a precipitous drop in objectivity up to $x = 4$ (2). Then there is a pushback in which sentiments cool off up to $x = 11$ (3). Finally, the sentiments slowly return to baseline from $x > 11$ (4). Positivity and negativity follow each other and by definition mirror the trends in objectivity.

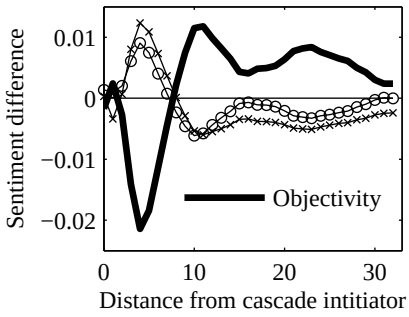
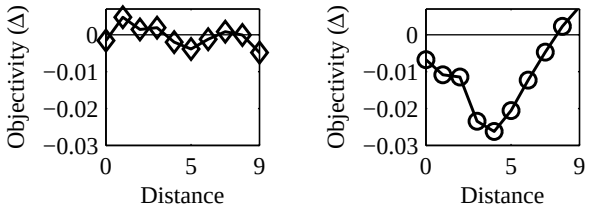


Figure 3: Change in relative-to-baseline objectivity (dark line), positivity (line with circle markers), and negativity (line with x markers) with respect to minimum path distance to cascade initiator. The max width of the 95% confidence interval over this distance range is 3.6e-4, which is too small to see in the figure.

The trends apparent from the emoticon-based approach are similar: a dramatic initial period with a peak, quick cool-off, and reversion to baseline. The comparable trend to that of sentiment derived from words suggests that emoticons provide a rough heuristic for the magnitude of a post's objectivity. However, the trends resulting from emoticon- and word-based approaches have opposite signs. A possible explanation is that emoticons are used to temper or hedge strong sentiment, such as "We really need that paper by tomorrow or our boss is going to kill us :) ".

Do all cascades have such a dramatic adolescence? To answer this question we define the highly variable initial phase as being made up of nodes whose shortest path distance is



(a) Shallow cascades (b) Deep cascades

Figure 4: Change of objectivity over baseline as a function of distance from the cascade initiator for shallow cascades (max $L < 9$) and deep cascades (max $L >= 9$).

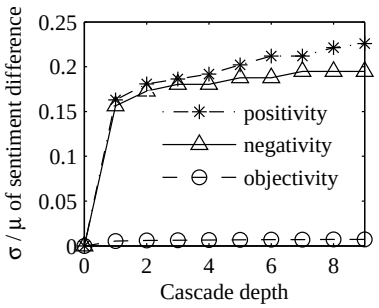


Figure 5: Coefficient of variation (σ/μ) of relative-to-baseline positivity, negativity, and objectivity. The values are calculated within nodes of a particular cascade at a given cascade depth, L, averaged over all cascades.

< 9, and the less dramatic tail as nodes with distance ≥ 9 nodes. Figure 4(a) shows the objectivity only for deep cascades (max $L \geq 9$) and Figure 4(b) shows objectivity just for shallow cascades (max $L < 9$). This plot strongly suggests that only deep cascades experience a dramatic adolescence.

Shallow cascades exhibit mundane use of subjective language. They start off slightly supportive of the initial post, using language with less negativity than usual, and then quickly peter out. A likely explanation is that these posts tend to be relatively tame, and so do not attract the attention of more posters. Thus, the cascade dies quickly. Note that at $x = 1$, where the first responders in short cascades are more supportive of the initial post, there is a corresponding blip of increased objectivity in Figure 3 for all cascades. The separate plots of deep and shallow cascades in Figure 4 reveal that deep cascades do not show this blip of increased objectivity. Thus, the increase at $x = 1$ in the objectivity plot over all cascades is caused by the overwhelmingly high count of short cascades with a characteristically supportive first responder.

In contrast to the humdrum sentiment of posts in shallow cascades, posts in deep cascades immediately use more heated language than usual and then rapidly cool off. This dramatic outburst may inspire other authors to comment, thus creating deep cascades. After this rush of emotions, deep cascades slowly revert to baseline as expected from the parent-child moderation previously discussed.

Previous work such as Adamic and Glance (2005) suggests that responders may form separate factions with polarized emotion. This is not in contrast to the trend of emotion returning to baseline; rather, emotion returns to baseline within these polarized factions. Figure 5 shows that sentiment does indeed polarize over time. At a given depth level in a cascade, there is a higher coefficient of variation for positivity and negativity than for objectivity. Our exploration into the relationship between variation at a given level in one cascade and tree distance between nodes yielded no clear trends.

Our results indicate that at the cascade level, the tendency to return to baseline is modulated by the cascade topology, specifically a post's position in the cascade and the overall depth of the cascade.

Conclusion

This paper traces the flow of sentiment through the blogosphere. We discover that the sentiment of a blog post is affected not only by the sentiment of its immediate parent, but also by its position within a cascade and that cascade's characteristics. We examine both the degree of sentiment (objectivity/subjectivity) and its polarity (positivity/negativity). Our analysis yields several conclusions: (1) nodes are strongly influenced by their immediate neighbors, (2) deep cascades show four distinct phases in their objectivity: typical, rapid heating up, rapid cooling off, and slow return to baseline, (3) shallow cascades, in contrast, have a brief and mild supportive tendency and then tend to remain as objective as normal, (4) sentiment polarizes within a cascade as depth increases, and (5) emoticon tagging is a rough heuristic for post sentiment, but augmented bag-of-words lexicon-tagged models provide a much more nuanced understanding of sentiment.

We believe that the idea of combining massive linguistically-tagged datasets with topological analysis offers many fruitful areas for further investigation.

Acknowledgments

We would like Stanford University for resources. This work was also supported in part by the National Science Foundation Graduate Research Fellowship and ONR grant No. N00014-10-1-0109.

References

Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 U.S. election. In *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05*, 36–43.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2200—2204.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, 497.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; and Glance, N. 2007. Cascading behavior in large blog graphs: Patterns and a model. In *SIAM International Conference on Data Mining*.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, 79–86.

Stone, P.; Dunphy, D.; Smith, M.; and Ogilvie, D. 1966. *The General Inquirer: A Computer Approach to Content Analysis.* The MIT Press.

Zafarani, R.; Cole, W.; and Liu, H. 2010. Sentiment propagation in social networks: A case study in LiveJournal. In *Advances in Social Computing*, volume 6007 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg. 413–420.