Sentiment Flow Through Hyperlink Networks

Mahalia Miller
Stanford University
Stanford, CA, USA.
mahalia@stanford.edu

Conal Sathi
Stanford University
Stanford, CA, USA.
csathi@stanford.edu

Daniel Wiesenthal
Stanford University
Stanford, CA, USA.
jepense@stanford.edu

Abstract

This project explores the flow of sentiment information through a network of hyperlinked web pages. We analyze the main text in a large set of web pages according to several rough heuristics for sentiment information. We then model the data as a network in which nodes represent pages and (directed) edges represent hyperlinks between pages. Within cascades we identify in the graph, we examine how the sentiment features of a node change as a function of its position in the cascade structure and the sentiment features of its cascade brethren, essentially capturing the flow of sentiment through the cascade. As well as analyzing the flow of individual sentiment features, we explore the interaction between objectivity and the other features. Our analysis yields several interesting conclusions: blog posts in which the author uses more objective language than usual are more likely to be cascade initiators, long cascades show qualitatively different behavior than short cascades, we see different behavior for "first round" and "second round" contributors to discussions, and subjective language leads to more subjective language at an increasing rate.

1 Credits

This paper is based off work for Jure Leskovec's Social and Information Networks CS 224w course and Dan Jurafsky and Christopher Potts Extracting Social Meaning and Sentiment CS424p course at Stanford University, Autumn 2010.

2 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 the interactions between authors and the trends emergent in the language they use. News articles and blogs 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 blog post. In this way, blog authors create a directed graph in which blog posts represent nodes and hyperlinks represent edges. Previous work on both sentiment and network analysis inspire several interesting areas for exploration in the overlap of the two mostly disparate fields. Is the sentiment of blog post cascade initiators unusual or predictable? How does sentiment flow through a cascade? Are there noticeable differences in the sentiment characteristics of blogs in long cascades versus short cascades? How do different sentimental aspects interact across the network? Do certain sentiment properties modulate the flow of other properties? In this paper, we propose a preliminary approach to answering these questions, apply our approach to a large data set, and present significant results showing clear trends with real-world implications.

3 Prior Work

Prior work is split in two directions: understanding the network characteristics of the diffusion of information in static and time-varying graphs, and classifying sentiment in the text of actual posts in isolation.

3.1 Prior Work on Cascades

Recently researchers have analyzed basic linking patterns in large political blog graphs. Adamis and Glance (2005) examine different linking trends in

Democratic and Republican blogs. They note a divided blogosphere where liberal and conservative blogs tend to link to their own communities, rather than to each other. They also discover the conservative blogosphere to be more densely connected and that the two communities tend to link to different news sources. There has also been work on analyzing patterns of cascading behavior in blog graphs. Lescovec et al (2007) discover temporal patterns and topological patterns. They note common cascade shapes, such as stars and chains, and cascade topological properties, such as the size of cascades and their degree distribution. We will examine topological features when assessing how basic topology, such as cascade shape and size, affects network flow of sentiment.

3.2 Previous Work on Sentiment Extraction and Propagation of Sentiments

One interesting challenge in sentiment analysis is identifying the presence of opinion in text and, if such an opinion exists, the polarity of the opinion. With the rise of Internet blogs and selfpublishing, it is becoming increasingly easy for a user to post information to either strictly inform others or to express a personal opinion on an issue. Researchers have tried various approaches. There has been research in using bag-of-word models and sentiment lexicons. Additionally, Chesley et al. (2006) employ verb classes and Wiktionary to get definitions and synonyms for adjectives. Godbole et al. (2007) use WordNet distances, and Zafarni et al. (2010) employ Normalized Google Distance. The two fields of information propagation/diffusion and sentiment analysis seem to be largely disparate, inviting research into their marriage. The authors are unaware of previous work tracking extracted sentiment propagation in a prose network. Closely related is the work of Zafarani et al. (2010), which examines data from Live Journal, a social network where users maintain a blog, and in which users have the option of assigning a mood to each of their posts. The paper explores tracking this manually assigned sentiment across the network by assigning each mood to a score and by defining mathematically whether the sentiment of one user has propagated to another. Our work departs the seminal ideas in Zarafani et al. (2010) in that we use links between blog posts, rather than friendships on a social network to denote edges, and we examine blogs from multiple sources. Further, we use the content of the post to extract sentiment ourselves, rather than relying on the user to explicitly describe his sentiment. This approach is subject to the limitations of our sentiment extraction, which we address by verifying results across two independent sentiment lexica. 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.

4 Methodology

4.1 Data Set

We obtained data from Jure Leskovecs memetracker project (Leskovec et al., 2009) from the month of August 2010, which consists of roughly 1 million blog posts per day. 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.

4.2 Pruning of the Data Set

Our data set initially comprised nearly 200GB of highly compressed data representing a full month of blogs and web pages, which allowed the luxury of pruning out unwieldy parts of the data set without encountering serious sparsity issues. We first examine the 15 top domains in our data, as shown in Table 1. Much of the original data stems from non-English domains (.fr, .jp, etc), which we trim out since our lexica are only applicable to English-language posts. As shown in Figure 1, Facebook and Twitter dominate the data set, comprising 66% of the data. These posts tend to be short, informal, and linguistically noisy. Also, most of these posts tend to be singletons (i.e. they tend to not cite other posts), so they will not be useful in our cascade analysis. Figure 1 shows the degree distribution of nodes in our network with and without Facebook/Twitter. The CDF without Facebook and Twitter shows a more gradual increase, indicating that a large number of Facebook and Twitter posts have a very low degree. After pruning

Blog	% of Posts in Network
twitter.com	41.59 %
facebook.com	24.63 %
blip.tv	2.98 %
t.love.com	0.68 %
blog.sina.com.cn	0.46 %
feeds.feedblitz.com	0.40 %
zoom.nl	0.24 %
islive.nl	0.16 %
plaza.rakuten.co.jp	0.14 %
capi-logement-neuf.fr	0.13 %
photobucket.com	0.11 %
bizjournals.com	0.11 %
snowboard-online-shop.eu	0.10 %
blogs.msdn.com	0.09 %
article.wn.com	0.07 %

Table 1: Top 15 Blogs.

Facebook, Twitter, and non-English root domains, we further prune singleton nodes (nodes with 0 in- and out-degree) from our dataset. This yields a large, linguistically rich, highly connected graph of roughly 8 million blog posts and 15 million hyperlinked edges.[1]

4.3 Sentiment Extraction

We extract the sentiment from each post by cross-referencing the words in the post (treated without order, as a bag-of-words model) against two sentiment lexica, and tracking frequency, counts, and averages for positive and negative scores, as well as a synthesized objectivity score. The two sentiment lexicons we use are the Harvard Inquirer (Stone et al., 1966) and SentiWordNet 3 (Stefano et al., 2010). By using two separate lexica which both purport to give information about positivity and negativity, we are able to compare and validate our results in cases where the two lexicons agreed strongly. In the Harvard Inquirer, each word is given a binary score (1 or 0) for positivity, and separately for negativity. In SentiWordNet, each word is assigned a score on a continuous scale between 0 and 1 for positivity and negativity. To

[1]We had to use efficient coding techniques to manage a data set of this size. Calculating sentiment features took over 24 hours distributed over eight 3+ GHz processor cores. Constructing the graph took 20+ hours. Since we could never hold our whole data set in RAM, we had to serialize to disk in between code segments.

Blog	Average Word Negativity
raphaelku.livejournal.com	9.9e-4
vadim-i-z.livejournal.com	9.9e-4
darkgothiclolita .forumcommunity.net	1.6e-2

Table 2: Sample negativity sentiment values.

synthesize objectivity, we assign a word a value of $1 - (SWNpositivity + SWNnegativity)$, as suggested by the SentiWordNet authors. Thus, for each post, we count and average these word scores to compute the objectivity/subjectivity and positivity/negativity of each post. We average these scores across all posts for each web domain (i.e. for nytimes.com, techcrunch.com, etc.) to create a baseline value. This baseline represents the authors usual manner of writing; since global averages do not account for differing writing styles, we find it more intuitive to compare against this local baseline in order to determine whether a particular post is unusual for its author. We present a motivating example below in Table 2, in which the third domains average negativity is more than 15 times the average negativity of the first two domains. Thus, for each post, we calculate all sentiment features as a delta above its baseline. In addition to creating a baseline for each node, we calculate a global mean, median, and standard deviation for each feature to get a sense of our data (the suffix ".Inq" denotes the feature came from calculations using the HarvardInquirer while the suffix ".swn" denotes the feature came from SentiWordNet) as shown in Table3. The average word ob-

Word Feature	Median	Mean	SD
posFreq.Inq	0	0.00915	0.0187
negFreq.Inq	0	0.00493	0.0147
emoFreq.Inq	0.00151	0.01408	0.0246
avgPos.swn	0.0139	0.01794	0.0192
avgNeg.swn	0.00973	0.0145	0.0178
avgObj.swn	0.973	0.967	0.0315
totalWords	105	333	1147

Table 3: Statistics for sentiment values.

jectivity has the highest standard deviation, which makes sense given the fact that objectivity is affected by both positivity and negativity, and so has more variability.

4.4 Cascade Identification

We construct a network graph of our data using the C++ SNAP network analysis library by Jure Leskovec. We model 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 are specifically interested in information propagation graphs (cascades). After constructing the graph, we search for "cascade initiators" by iterating through each node and looked for nodes with a non-zero out-degree and a zero in-degree. These nodes represent posts that begin a long chain of links. Once cascade initiators are found, we search for all the nodes in each cascade by applying a breadth-first algorithm, starting at each cascade initiator and following the in-links. A small example cascade of depth 2 is shown in Figure 2. In this example, the cascade initiator is the node on the left.

5 Analysis

5.1 Global Analysis

Our primary analysis focuses on the change over baseline of sentiment features for a post as the shortest-path distance to the post from the cascade initiator increases. We see an interesting trend in the average positivity and negativity values for nodes at distance d away from the cascade initiator, shown in Figure 3. We note several qualities of this plot. The cascade initiators tend to have lower average positivity and negativity values, suggesting (as we explore in more depth below) that posts that initiate cascades tend to be written in a more objective manner than is usual for the writer. The first two posters after the initiator also tend to have lower average positivity and negativity (relative to their respective baselines), but for nodes at a distance between 4 and 9 we see the average positivity and negativity diverge to mirror each other across the baseline (x axis), with an increase in average negativity. After a distance of 9, most nodes are slightly below baseline for both features. As distance increases, the two features begin to diverge from baseline, and after node 200 our data becomes sparse so outliers destroy the trends. We suggest the following explanation for these observations: posts that are more objective than usual tend to have more interesting factual content for readers to discuss (for example, an original article on networks). As such, they tend to be largely objective and generate cascades with greater longevity. The first few posters after the initiator contribute interesting ideas, hence their low pos/neg frequencies relative to baseline ("You make an interesting point here, which might apply to the field of NLP"). After this first round of observers arrives on the scene, the second round of commentators arrives, with a set of writers and opinions laid out before them: they have more to criticize, both in terms of raw material and in terms of the interactions between their predecessors ("You completely failed to understand the original article, think before you write, you $< insult >$!").

We note that the initially high objectivity, which then drops and picks up again, may be an artifact of two qualitatively different types of cascades: long and short. That is, cascades that are hundreds of posts long may behave in some qualitatively different manner than their shorter brethren. We break down our data to compare not only the whole network graph, but also variants where we consider only nodes participating in cascades shorter than 10 nodes in depth or greater than 10 nodes in depth, shown in Figure 4.

We see that indeed, our above observations hold for the large cascades, and are even intensified. There is a qualitatively different trend across small cascades: the cascade initiators have significantly higher average positivity and negativity than baseline, but their responders immediately have lower than baseline values for these metrics. We hypothesize that for these short cascades, the initial post contains incitatory language that prompts a few responders to feel the need to clarify by "laying out the facts," or to calm the author down. As average positivity/negativity does not necessarily allow us to measure objectivity, we also examine objectivity for the whole network, as well as the small- and large-cascade subsets. This is shown in Figure 5. We find that, consistent with our above findings, objectivity is far above baseline at the node initiator, and decreases slightly for nodes at a distance between 4 and 9 away from the cascade initiator, though it remains above baseline for most of the cascade, only flaring out at the ends of long cascades. These findings are consistent when we examine both the small- and large-cascade subsets of our graph. Thus, over both large and small cascades, we find that objectivity stays above baseline except at the very end of the cascade. We hypothesize that the flow of causality here is from above-average objectivity towards longer cascades: that is, more objective writing inspires cascades, while peaks of heated rhetoric and strong sentiment quickly kill cascades.

5.2 Local Analysis

In addition to examining the sentiment values at a scale relative to their global position in a cascade, we examine sentiment values in the local context of parent-child node relationships. Figure 6 depicts the positivity of a post by its distance from the cascade initiator, taking the objectivity of its parent into account. The blue represents nodes whose parents have an objectivity value of 0.99 (i.e. the node cites a very objective post), whereas the red represents nodes whose parents have an objectivity value of 0.89 (i.e. the node cites a very subjective post). The rest of the color spectrum is consistent in the plot: that is, as a node's parent becomes more subjective, the color on the plot becomes warmer. We see that as the subjectivity of a node's parent increases, the positivity of the node becomes harder to predict (as indicated by the much lower correlation coefficient for the nodes with more subjective parents). We offer the following explanation: a more opinionated post will elicit a more powerful emotional response. (Contrast this to a very objective post, which allows agreement and disagreement, but is not as likely to inspire an impassioned response.) Within this response, there will be posts that strongly agree and nodes that strongly disagree. It is very hard to predict the polarity of the response, though the data show that a high-arousal response is common. This relationship of parent-to-child objectivity merits further examination. We see in Figure 7 that if a node's parent is unusually subjective, it is more likely that the node will also be subjective. Intuitively, this might be explained as follows. If there is a post that is very subjective, it is very opinionated, thus, eliciting either strong agreement or strong disagreement. Thus, any post that cites this first post will express much sentiment in expressing this agreement or disagreement to the prior post's stance. Moreover, the slope of the graph gets progressively steeper. Thus, subjective language elicits more subjective language at an increasing rate. This may be the cause of flame wars in long cascades despite the cascade initiator itself being unusually objective.

6 Conclusion

We discover that the sentiment of a blog post is affected by its position in a cascade as well as by the objectivity of its immediate parent. We examine both the degree of sentiment (objectivity/subjectivity) as well as explore the polarity (positivity/negativity). Our analysis yields several interesting conclusions: (1) blog posts in which the author uses more objective language than usual are more likely to be cascade initiators (especially so for longer cascades), (2) long cascades show qualitatively different behavior than short cascades, with long cascades beginning with unusually objective posts and short cascades beginning with unusually subjective posts that are quickly responded to objectively, (3) we see different behavior for "first round" and "second round" contributors to discussions with the former offering more objective feedback and the latter using unusually subjective language (perhaps judgment), and (4) subjective language leads to more subjective language at an increasing rate. Moreover, our results hold using the corresponding sentiment scores from either the Harvard Inquirer or SentiWordNet showing that our results are not an artifact of our lexicon.

6.1 Future Work

Our work is one of the first research projects we know of that explores sentiment extraction and analysis across a large-scale multi-domain network. Naturally, there are many avenues to pursue for future work.

One avenue for further work is to build a more refined sentiment extraction engine. Use of a sentence parser to tag the parts of speech of words in the post would help disambiguate multiple meanings of a word to better extract its sentiment. For example, our sentiment lexica currently score words like "well" with high positivity and low negativity, unaware of whether the word is a noun or adverb–yet these two parts of speech carry significantly different sentiment information. Thus, finding the part of speech of the word could help to disambiguate the sense of the word. Similarly, a full sentence tree parse would help to correctly classify special cases where a word's presence is not necessarily an indication of its effect on the overall post sentiment, such as in the case of negation. (For example, when any negation (such as "no," "not," etc.) precedes a word, its sentiment should reverse, such as in the case "I am not feeling good about the outcome of this election.")

Often in this paper we discuss positivity as being related to agreement of a previous post's agreement and negativity as being related to disagreement, which is a simplifying assumption. It would be be interesting to actually see whether this is the case by looking at the content of each post and examining the language related to agreement/disagreement, in addition to positivity and negativity, to further refine our results. This might be done by starting with a seed set of words that strongly indicate agreement and disagreement (potentially these two words themselves) and then propagating through WordNet to expand the set, with agreement/disagreement scores decreasing with increase in distance from the seed set.

We would also like to explore more network graph properties in conjunction with our sentiment findings. We have largely focused on directed relationships and distance from the cascade initiator to a node on the cascade, but looking at the effect of cascade tree breadth versus depth is an area ripe with possibilities. For example, we might ask whether flame-wars have a tall and narrow graph structure, perhaps with many flips in change from baseline (a very positive post, then a very negative one), versus short, fat cascades where many secondary authors interpret a single seminal article in varying ways as they bridge it to their fields of interest, without necessarily being aware of their peers.

We would like to extend both our network and sentiment features to predict a node's sentiment values with a classifier. As our feature set grows more complex, it becomes difficult to manually sort out interactions between features, and a MaxEnt classifier may help us identify interesting interactions for close examination. The combination of a writer's history (baseline) and more nuanced hybrid semantic-network features (the sentiment of a friend who links to the same post, for example), could be a powerful technique for predicting a person's reaction to a given post. This has a plethora of applications: advertisement targeting, personalized news feeds, and political approval prediction, among others.

Acknowledgments

Mahalia Miller was partially funded by the National Science Foundation and the Stanford Graduate Fellowship. Daniel Wiesenthal and Conal Sathi were partially fueled by Starbucks.

References

Lada Adamic. and Natalie Glance. 2005. The political blogosphere and the 2004 U.S. election. *Workshop on Link Discovery*

Paula Chesley, Bruce Vincent, Li Xu, and Rohini Srihari. 2006. Using verbs and adjectives to automatically classify blog sentiment *AAAI Symposium on Computational Approaches to Analysing Weblogs*, 27–29.

Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. *Proceedings of the International Conference on Weblogs and Social Media*

Jure Leskovec, Mary McGlohen, Christos Faloutsos, Natalie Glance and Matthew Hurst. 2007. Cascading behavior in large blog graphs: Patterns and a model. *SIAM International Conference on Data Mining*

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the Dynamics of the News Cycle. *ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining.*

Jure Leskovec. 2010. *Dynamics of large networks. PhD Dissertation*. Machine Learning Department, Carnegie Mellon University, Technical report CMU-ML-08-111, 2008.

B. Stefano, A. Esuli, and F. Sebastani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*: 2200-2204.

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

Reza Zafarani, William Cole, and Huan Liu. 2010. Sentiment extraction in social networks: a case study in LiveJournal. *In Advances in Social Computing*, 6007:413–420.

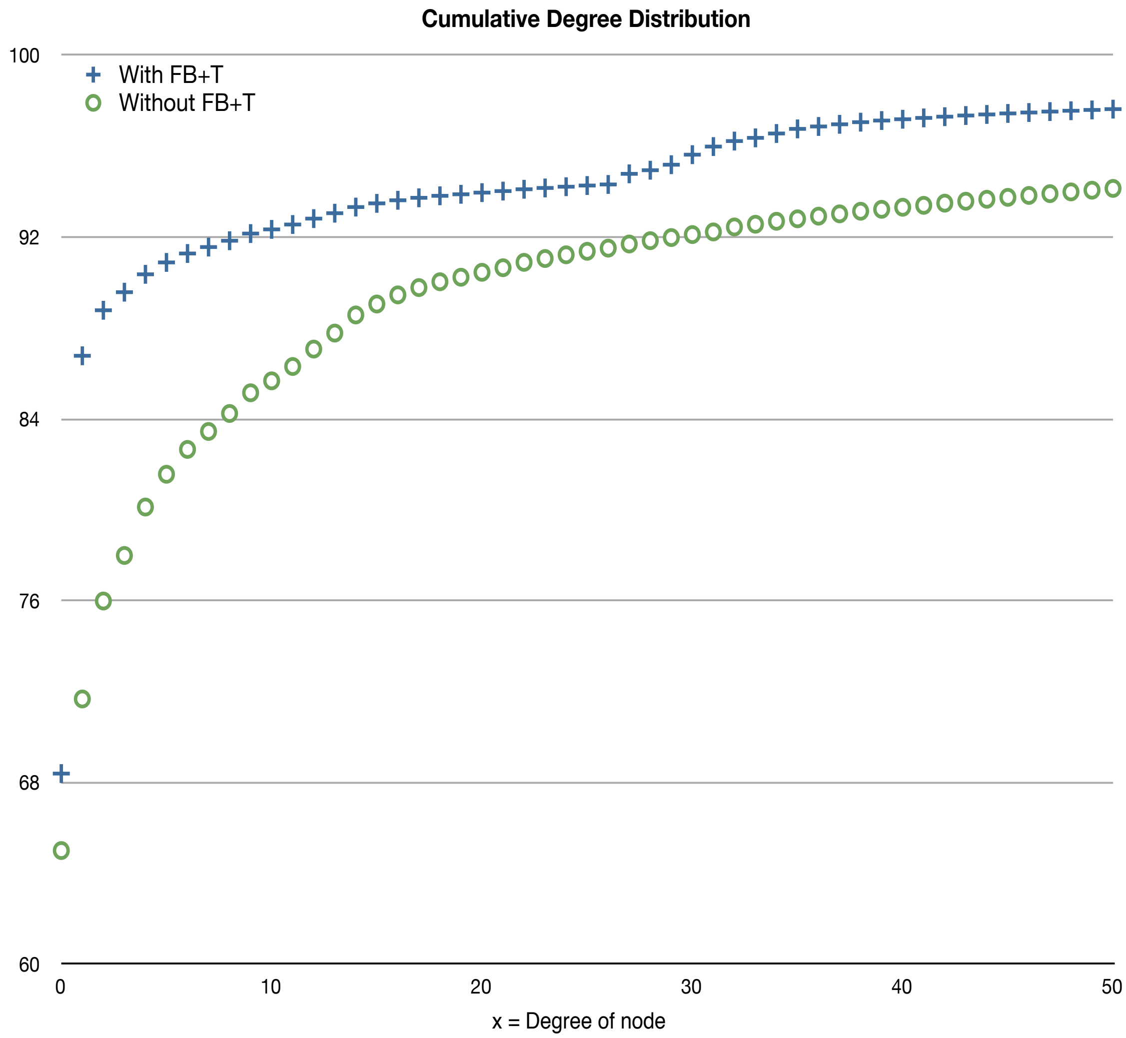

Figure 1: Cumulative degree distribution for node degree

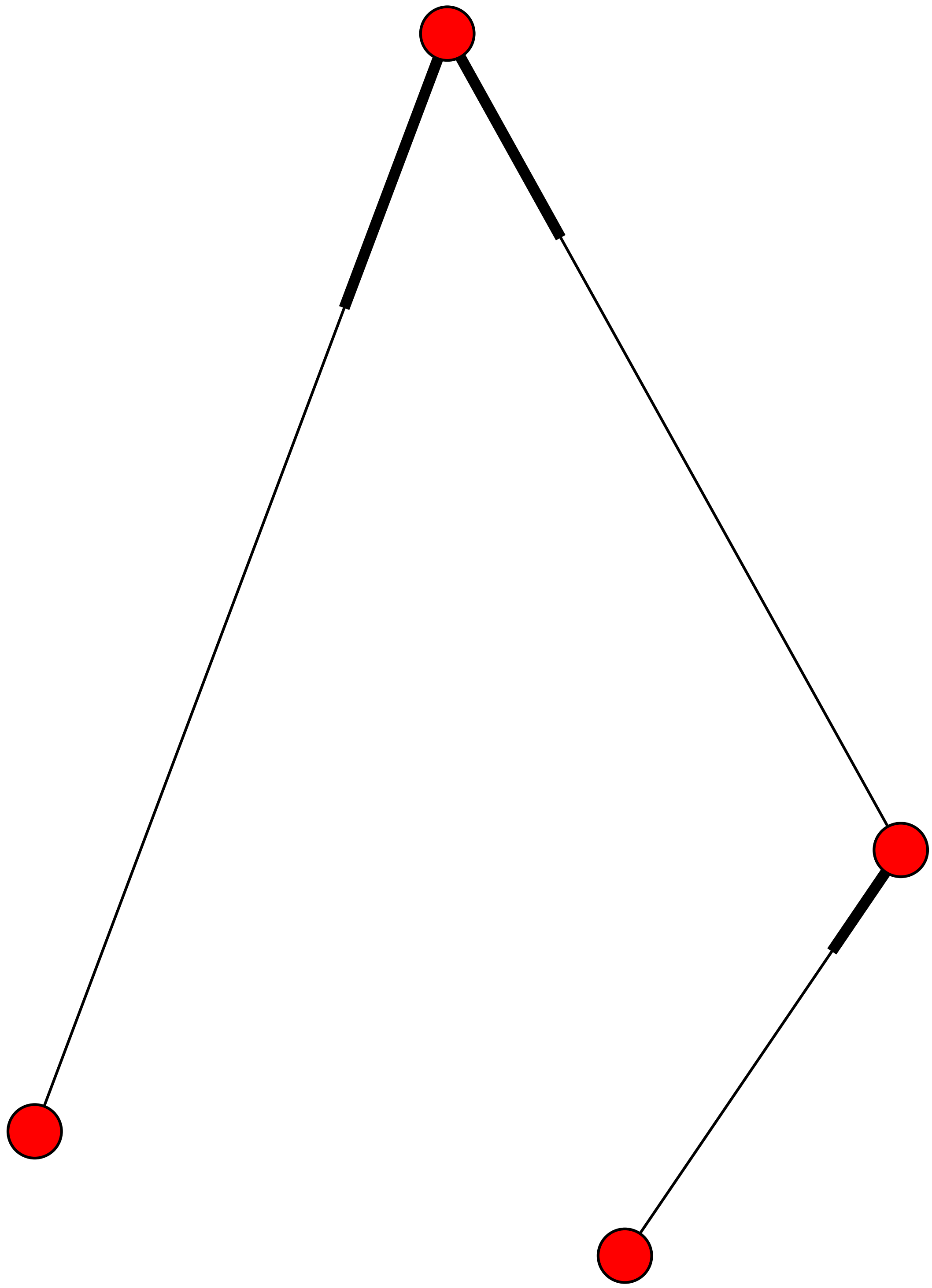

Figure 2: Sample cascade from data set

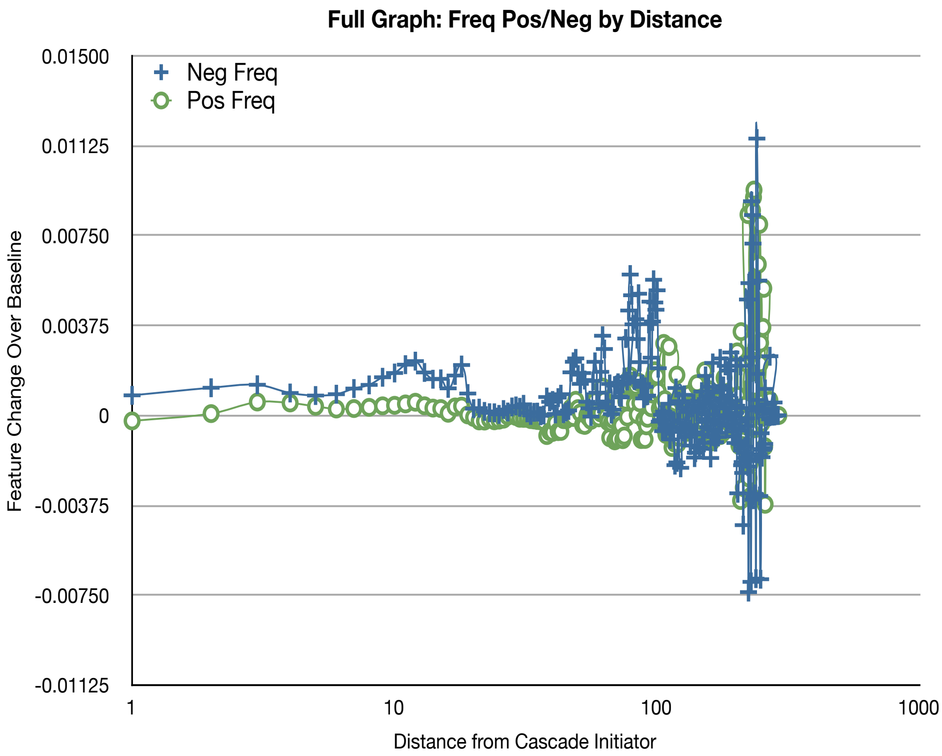

Figure 3: Average positive and negative word frequency for each node over each node's own baseline

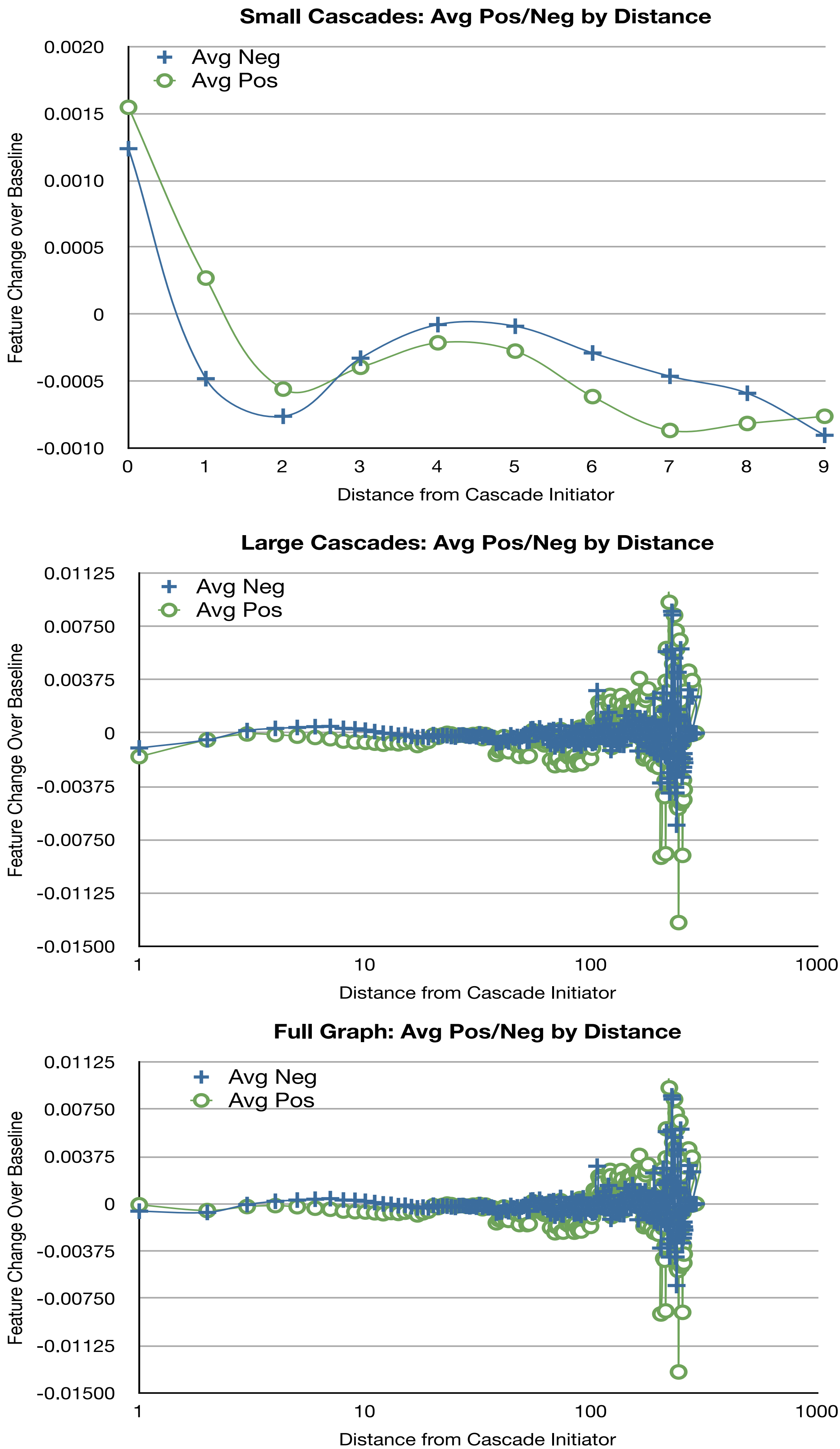

Figure 4: Average positivity and negativity for each node over each node's own baseline

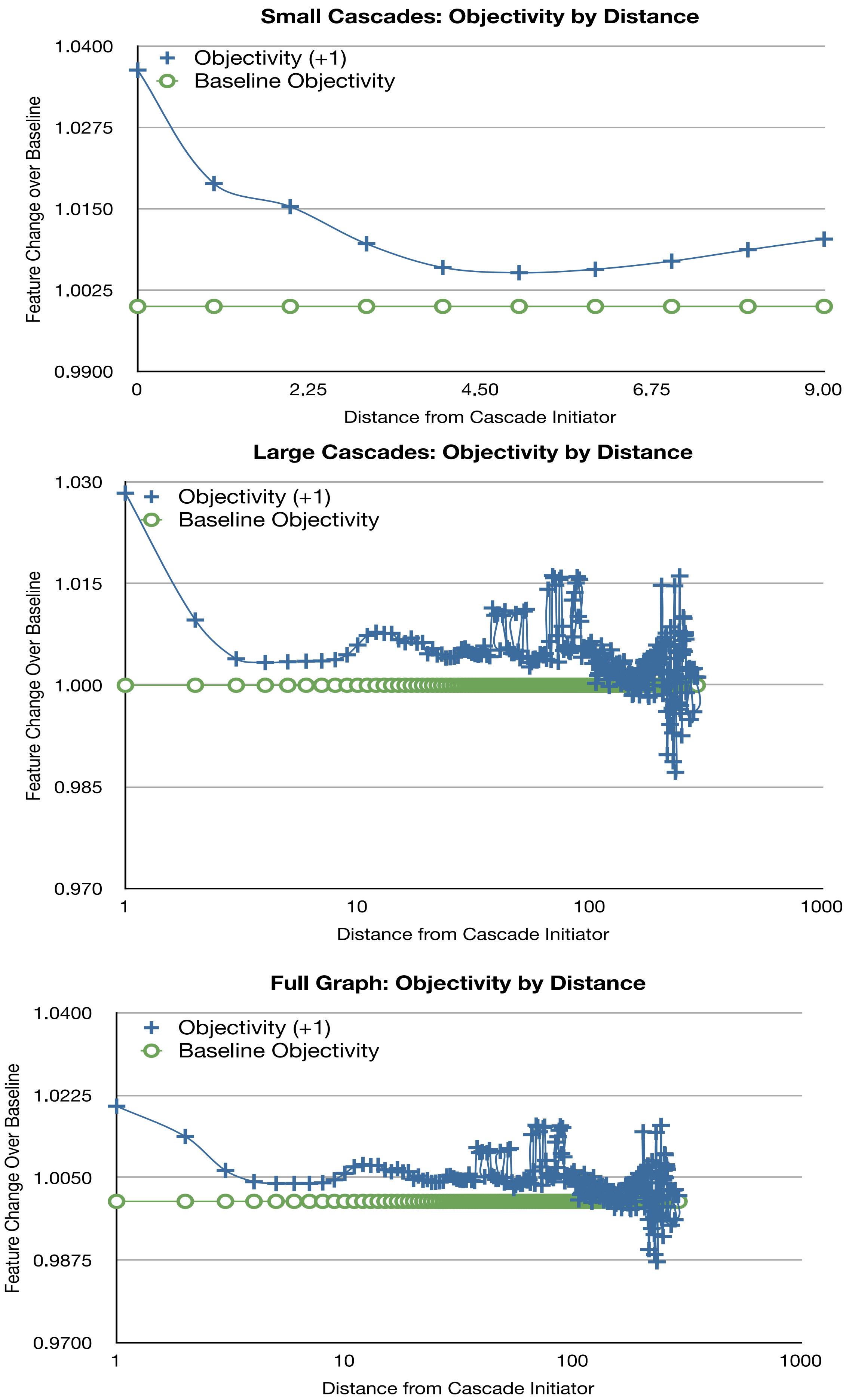

Figure 5: Objectivity for each node over each node’s own baseline

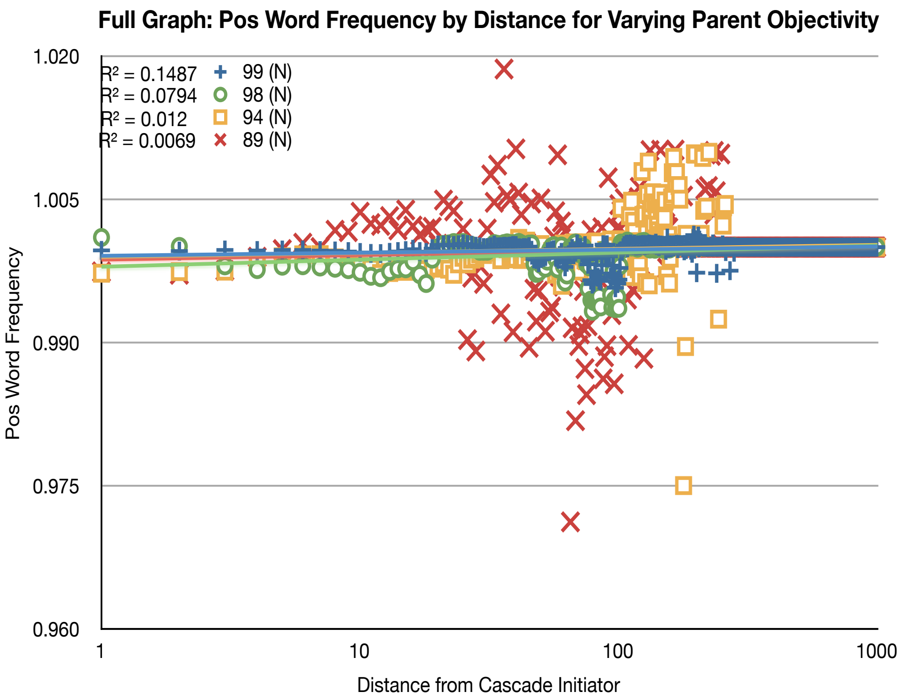

Figure 6: Average positivity for different levels of objectivity of the parent on the shortest path

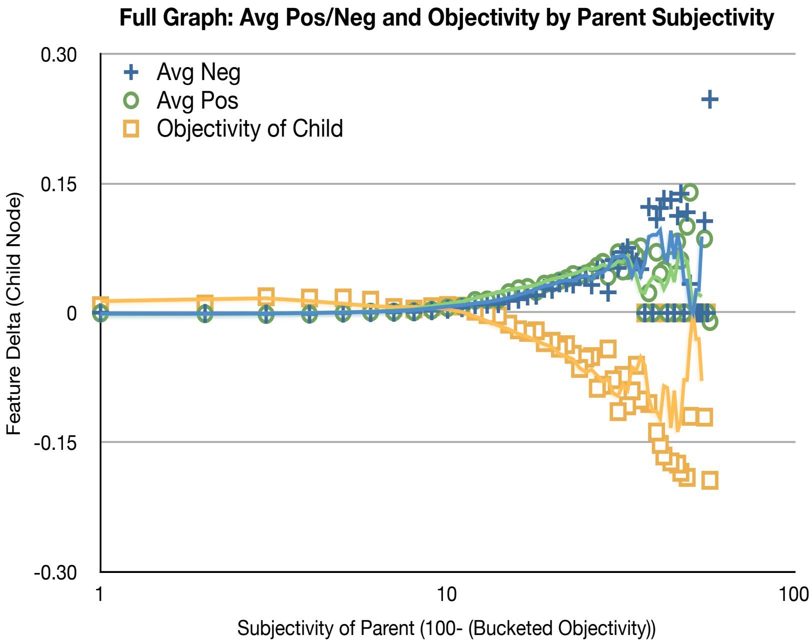

Figure 7: Average positivity, negativity, and objectivity versus the subjectivity of the parent on the shortest path