

# Modeling Trust and Influence in the Blogosphere Using Link Polarity<sup>1</sup>

Anubhav Kale, Amit Karandikar, Pranam Kolari, Akshay Java, Tim Finin, Anupam Joshi

University of Maryland, Baltimore County

Baltimore MD 21250, USA

{akale1, amikt1, kolari1, aks1, finin, joshi}@cs.umbc.edu

## ABSTRACT

The role of social networks has been well explored in understanding how communities and individuals spread influence. In a densely connected world where much of our communication happens online, social media and networks have a great potential in influencing our thoughts and actions. We describe techniques to find "like minded" blogs based on blog-to-blog link sentiment for a particular domain. Using simple sentiment detection techniques, we identify the polarity (positive, negative or neutral) of the text surrounding links that point from one blog post to another. We use trust propagation models to spread this sentiment from a subset of connected blogs to other blogs and deduce like-minded blogs in the blog graph. Our results confirm that the simple heuristics for analysis of text surrounding links and generation of missing polar links (links with positive or negative sentiment) using trust propagation is highly applicable for domains having weak link structure. These techniques demonstrate the potential of using polar links for more generic problems such as detecting trustworthy nodes in web graphs.

## Keywords

Blog, sentiment detection, link polarity, trust propagation

## 1. INTRODUCTION

Social media is a dynamic and growing area that includes collection of blogs, wikis, forums, photos and videos sharing sites. According to wikipedia<sup>1</sup> "social media describes the online tools and platforms that people use to share opinions, insights, experiences, and perspectives with each other". What makes social media really interesting is the level of user participation and conversations around different topics. This leads to formation of communities around topics that can vary from politics, technology, arts to knitting or photography and public relations.

Blogs have become a means by which new ideas and information spreads rapidly on the web. Blogs in general contain mostly user generated content, as do other sites like delicious<sup>2</sup>, flickr<sup>3</sup> etc. Bloggers link to interesting posts or might even comment on someone else's blog and these links tend to be the basis of conversations. As communities in social media like blogs emerge, there are thought leaders and individuals who have a significant share of the community's attention. *Influential* nodes in a social network can be responsible for starting a buzz or getting the community to notice a new trend or product. Monitoring and tracking both influential nodes and their opinions on the blogosphere, can thus have a significant number of applications in the realm of product marketing.

In this paper, we address the problem of modeling trust and influence in the blogosphere. Our approach uses the link structure of the blog graph to associate sentiments with the links connecting two blogs. (By "link" we mean the url that blogger *a* uses in his blog post to refer to blogger *b*'s post). We call this sentiment as *link polarity* and the sign and magnitude of this value is based on the sentiment of text surrounding the link. These polar edges indicate the *bias/trust/distrust* between the respective blogs. In order to associate a given blog *foo* to the community of its like-minded blogs, we *create* new *polar links* between all pairs of blogs using initial *polar links*. We use trust propagation models to "spread" the initial polarity values to all possible pairs of nodes. Finally, we compute the trust/distrust score for *foo* from the seed set of *influential* blogs (discussed later) to determine its community. More generally, we address the problem of detecting all such nodes that a given node would trust even if it is not directly connected to them.

Our approach uses simple shallow natural language processing to determine *link polarity*, yet preliminary results indicate that our approach has the potential to aid conventional community detection techniques based on path distance and reachability metrics. Since, we do not process entire blog-post text for sentiment detection between two blogs and use shallow NLP techniques, we speculate that the approach should scale well for real-time applications (e.g., analyzing blogs for dynamic situations like elections) than traditional off-line and computation intense approaches. This paper presents some of our preliminary results in the domain of blogosphere, however a long-term goal of our work is to deduce trustworthy nodes for a given node in any web-graph. We believe that *directed polar links* have a tremendous potential for addressing this hard problem.

## 1.1 Background

Bloggers typically discuss views about varied topics and are based on personal experiences. Such views are expressed almost instantaneously as soon as any new event occurs. The blogosphere has matured a lot since its inception and hence, when an event occurs, the first reaction is to turn to the blogosphere to see what people are saying about it. For example during the London bombings, people were interested in finding first hand reports, pictures, emotions and experiences of Londoners. As time progressed, people might have looked for more information about the event - what happened? Why? How many people were killed? Have there been any arrests? Which group(s) has claimed responsibility for this act? etc.

Suppose that your goal was to market a new kind of mp3 player which would compete with ipod. One of the starting points

---

<sup>1</sup> Partial support was provided by NSF awards ITR-IIS-0326460 and ITR-IDM-0219649 and funding from I.B.M.

---

1 <http://en.wikipedia.org/>

2 <http://del.icio.us>

3 <http://www.flickr.com>

would be to use advertising products such as Google's popular AdSense network. Using content of the webpage, this service matches relevant web pages to advertisements that relate to the topic of the page. While this gives a wide coverage and a significant audience, there is very little the advertiser can do to actively promote the product to the *right set of individuals*. Using a blog search engine one can find a ranked list of relevant blog posts for different generic query terms. However, most blog search engines use link based ranking schemes that measure popularity as opposed to influence. While a number of popular bloggers may talk about ipods in general, if the marketing division of your company can target the *community* that has a negative bias about ipod then chances of spreading good word about the new mp3 player is considerably high than targeting the *community* having a strong positive bias about ipod already. Thus having an insight into the communities in social media can aid in accurately targeting key personnel for marketing new products.

Temporal analysis of the swing in trends among communities has interesting applications for scenarios such as elections where a study of cause and effect phenomena has tremendous potential to gain an insight into change in voters' (or bloggers') bias during the election campaign events. This further implies that a community detection system capable of performing highly efficient real-time analysis of streaming data from social media can play a vital role for analyzing the effects of a candidate's meetings, speeches etc during election time.

There has been considerable amount of work in cluster formation and community detection on web graphs, however to our knowledge; none of the prior work involves using polarity of links as a parameter for the problem of community detection. Also, most of the well-known clustering algorithms like [19] are based on the analysis of link structure and do not work well for sparsely connected graphs. Our work is an initial step to address this problem. The remainder of the paper proceeds as follows. Section 2 covers related work. Section 3 describes the details of our approach, heuristic and data modeling. Section 4 covers the experiments and we discuss conclusions and future work in section 5 and 6.

## 2. RELATED WORK

We believe that the research in the area of information propagation was inspired by a large body of work in disease and epidemic propagation. As described in [2] this model applies well in the blogosphere where a blogger may have a certain level of interest in a topic and is thus *susceptible* to talking about it. By discussing the topic he/she may *infect* others and over time might *recover*. The authors use this approach in characterizing individuals into various phases of a topic in which they are more likely to become *infected*. They model individual propagation and use an expectation-maximization algorithm to predict the likelihood of a blogger linking to another blogger. They also study the different types of topics present in the dataset and describe an approach to categorize topics into subtopics. Certain topics are more *infectious* than others and spread through the social network of bloggers. Our work is focused more towards a scalable approach for community detection than behavioral models for topical analysis. Adar et al. [3] have proposed the use of URL citations to infer the dynamics of *information epidemics* in the blog-space. They also show that the PageRank algorithm finds authoritative blogs. A variation, called "iRank", is described to rank blogs based on their *informativeness*. In this scheme, each directed edge is assigned a weight function that gives importance

to URL citations which are closer in time. The edge weights are then normalized and PageRank computation follows. This weighted graph is called the *implicit information flow graph*. iRank makes use of the temporal nature of blogs by differentially weighing each citation in the graph by the time difference between when the blog mentions a URL and how soon it is referenced by other blogs.

Since bloggers are constantly keeping abreast of the latest news and often talk about new trends before they peak, recent research has focused on extracting opinions and identifying buzz from blogs [17]. Arun Qamra et al [21] developed a model that incorporates the content of blog entries, their time-stamps, and the community structure to extract the temporal discussions occurring within blogger communities. Gruhl et al. [5] have found strong correlation between spikes in blog mentions to Amazon sales ranks of certain books. More recently, Lloyd et al. [6] found similar trends for named entities in blog mentions and RSS news feeds.

Ravi et al. [7] have analyzed the word burst models [8] and community structure on the blogosphere [9] and they found a rapid increase in the size of connected component on the blogosphere. They argue that this trend is due to the increasing tendency of bloggers to comment about other blogs. Their results on the size of strongly connected components aid in our hypothesis that sentiment detection using links rather than analyzing the post text has potential for results with high precision-recall.

Blog hosting tools, search services and Web 2.0 sites such as Flickr and del.icio.us have popularized the use of tags. Tags provide a simple scheme that helps people organize and manage their data. Tags across all the users, collectively, are termed as *folksonomy*, a recent term used to describe this type of user-generated content. Tags are similar to keywords used in the META tag of HTML. Some researchers have studied the phenomenon of user-generated tags to evaluate effectiveness of tagging. Brooks and Montanez [10] present an analysis of the 250 most frequently used Technorati tags. Their study finds that tagging can be helpful for grouping related items together but does not perform as well as text clustering. Shen and Wu [11] treat tags as nodes and the presence of multiple tags for a document as a link between the tags. According to Shen, the network structure and properties of such a graph resemble that of a scale free network. However none of them consider sentiment around the tag which makes our work significantly different. Marlow [12] compares blogroll links and permalinks (URLs of specific blog post) as features to determine authority and influence on the blogosphere. The study suggests that permalink citations can approximate influence. Present blog search engines indeed use permalink citations or inlinks to a blog as a measure of authority.

A number of researchers have worked on the problem of propagating trust in a networked environment. Yu and Singh [13] propose a framework based on the assumptions of symmetric and transitive trust. There has been significant work in modeling eBay trust. Kamvar, Schlosser et al [14] have proposed a framework to assign a universal trust to a given node in the graph. However, in our environment trust propagation bootstraps with pair-wise opinion/trust, hence their model does not work well for our settings. Guha et al [1] in their paper titled *Propagation of trust and distrust* cover work related to trust propagation in multiple

disciplines and claim that their work appears to be first “to incorporate distrust in a computational trust propagation setting”. We found that their work was most complete and the trust propagation model suits well to our domain. Hence, our trust propagation approach is very similar to their work.

To the best of our knowledge, no prior work exists in the area of blogosphere to assign sentiments to links (what we term as *link polarity*) and use such polar links to find like-minded blogs in graph.

### 3. PROPOSED APPROACH

In this section, we describe our proposed approach and set the basis for experimental validations. We also provide some details on Guha’s trust propagation technique wherever appropriate.

#### 3.1 Link polarity

The term *Link Polarity* represents the opinion of the source blog about the destination blog. The sign of polarity (positive, negative or zero) represents whether the bias is for, against or neutral and the magnitude represents how strong or weak the bias is. In order to detect the sentiment based on links, we analyze section of text around the link in the source blog post to determine the sentiment of source blogger about the destination blogger. From our analysis of blog texts and interactions with regular bloggers, we observed that it is not necessary to analyze the complete blog post text to determine the sentiment. In fact, text neighboring the link provides direct meaningful insight into blogger A’s opinion about blogger B. Hence, we consider a window of  $x$  characters ( $x$  is variable parameter for our experimental validations) before and after the link. Note that this set of  $2x$  characters does not include html tags.

#### 3.2 Sentiment detection

There has been considerable work on sentiment detection on free-form text. Researchers have experimented with various natural language processing techniques. However, we do not need to employ any complex natural language processing techniques since bloggers typically convey their *bias* about the post/blog pointed by the link using fairly standard vocabulary. Hence, we use a corpus of positive and negative oriented words and match the token words from the set of  $2x$  characters against this corpus to determine the polarity.

Since our corpus includes words in noun forms, it is essential for us to employ stemming on tokens. We apply stemming mechanism on all such tokens and then convert them into canonical form by eliminating characters such as commas, periods, exclamation marks etc. We observed that bloggers frequently use negation of sentimental words while indicating bias about another blog-post (“*What b says is not bad*”), hence our corpus also includes basic bi-grams of the form “not <positive/negative word>”. Our experiments confirmed that the aforementioned simple techniques are very effective in deducing the text sentiment correctly.

##### 3.2.1 Calculation of link polarity

The number of positive and negative words varies to a great extent (typically from 1 to 30 in window size of 750 characters) across multiple posts. Hence, it is necessary to normalize the results over some metric. We adopted the following formula for calculating the link polarity:

$$Polarity = ( Np - Nn ) / ( Np + Nn )$$

$Np$  : Number of positively oriented words

$Nn$  : Number of negatively oriented words

Notice that our formula incorporates zero polarity links automatically.

### 3.3 Trust Propagation

Since blog graphs are not densely connected, we still do not have the trust scores between any given pair of nodes. Hence, we must employ some *sentiment spread* mechanism to calculate trust score between all pairs of nodes from the set of nodes having polar edges between them. There has been considerable amount of work in computer science as well as other disciplines on various aspects of trust definitions, trust metrics, trust propagation models and validation techniques. Guha et al [1] have proposed a framework to spread trust in a network bootstrapped by a known set of trusted nodes. They have evaluated their approach on a large dataset from epinions<sup>4</sup>. Guha’s approach uses a “belief matrix” to represent the initial set of beliefs in the graph. This matrix is generated through a combination of known trust and distrust among a subset of nodes. This matrix is then iteratively modified by using “atomic propagations”. Finally “rounding” technique is applied on the matrix thus generated so far, to produce absolute values of trust (yes or no) between all pair of nodes. The “atomic propagation” step incorporates direct propagation, co-citation, transpose trust and trust coupling. The overall trust propagation mechanism is represented using matrix operations (additions, multiplications and transpose). We adapt this approach with some modifications for our work. The section on experiments covers our modifications in greater details.

In order to form clusters after the step of trust propagation, we take the approach of averaging trust score for all blog nodes from a predefined set of “trusted” nodes belonging to each community. A positive trust score indicates that the blog node belongs to the community *influenced* by the trusted node of that community. Specifically, we selected top three *influential* democratic and republican bloggers. (We address our notion of *influential* blogs shortly). A positive trust score for a blog *foo* from top three democratic blogs indicates that *foo* belongs to the democratic cluster and a negative score indicates that *foo* is a republican blogger. Notice that negative links thus help us to classify a blog into the right cluster even if it is not very well connected within its cluster. In order to determine the *influential* bloggers in each community we experimented with the heuristics of high incoming-degree, high outgoing degree and random subset of all nodes.

## 4. EXPERIMENTS

We now present the results of our experiments that demonstrate the feasibility and effectiveness of link polarity. Also, we describe the motivation behind choosing the political domain for our experiments and present a representative set of link polarity computations for some of the *influential* blogs.

### 4.1 Choice of domain

We decided to choose political blogs as our domain; one of the major goals of the experiments was to validate that our proposed approach can correctly classify the blogs into two sets: republican and democratic.

Through some manual analysis of the political blogs, we observed that the link density among political blogs is reasonably high and hence we could deduce the effectiveness of our approach by running our algorithms over fairly small number of blogs. In other words, we do not need to perform a large number of iterations of Guha’s atomic propagations; about 20 iterations suffice to *create* polar links with sufficiently accurate polarity values between blogs that did not link to each other.

The dataset from Buzzmetrics [15] provides link structure between blog posts over 1.3 million blog posts. Hence, we needed to aggregate this post-post link structure to a blog-blog link structure. This implied that we should choose such a domain where there would be minimal number of off-the-topic posts from the same blog and political blogs fit this requirement perfectly. (We address this issue of determining link polarity based on specific topics in our discussion section).

From a business model point of view, political blogs are highly effective during election period to determine the trends among voters and a technique that can classify voters into multiple political biases would be extremely beneficial to various sources.

### 4.2 Parameters for trust propagation

Guha’s work argues that “one step distrust” provides the best trust propagation results in their domain of experiments. They propose the notion of “trust and distrust” between two nodes in the graph where the same set of two nodes can trust or distrust each other. “one step distrust” uses “trust matrix” as the belief matrix. However, we believe that in our domain the initial belief matrix should incorporate both trust and distrust (positive and negative polarities from blog A to blog B). Hence, we use the difference between trust and distrust matrices as our initial belief matrix. We believe that the idea of using “eigenvalue propagation” to determine final trust scores is generic and applies to any domain. Hence we used the same for our experiments.

We experimented with various values of the “alpha vector” to confirm that Guha’s conclusion of using the values they proposed {0.4, 0.4, 0.1, 0.1} yields best results. Our experiments indeed confirmed that this set of values yield the most accurate results. We do not provide the results of our comparisons here, since this is not the contribution or the primary motivation of our work. Further, Guha et al recommend performing “atomic propagations” approximately 20 times to get best results. Since, we can not guarantee that such numbers would work in our domain; we took the approach of iteratively applying atomic propagations till convergence. Our experiments indeed indicate a value close to 20, after which the final trust scores do not seem to improve. Finally, we do not incorporate the extra step of “rounding” in Guha’s work since the sign of trust is sufficient to determine if the blog under consideration belongs to democratic or republican set.

### 4.3 Parameters for link polarity

As explained in section 3, we used various window sizes around the links to fetch the token words to be used for sentiment detection. After some manual analysis of political blogs, we decided to experiment with 1000, 750, 500, 250 and 50 characters before and after the link under consideration. We expected to get some insights into what would be the right window size (and

hence, the right number of words around links that yield more signal than noise) by varying this parameter.

### 4.4 Datasets

We studied the effectiveness of our approach over a graph of 300 blogs created from the link structure of buzzmetrics [15] dataset. We observed that in-degree as a heuristic works better over out-degree and random heuristics for selection of *influential* nodes for the seed set. Hence all the results that follow are based on the in-degree heuristic. Lada A. Adamic provided us with a reference dataset of 1490 blogs with a label of democratic or republican for each blog. Their data on political leaning is based on analysis of blog directories. Some blogs were labeled manually, based on incoming and outgoing links and posts around the time of the 2004 presidential election. Buzzmetrics does not provide a classified set of political blogs. Hence, for our experiments we used a snapshot of Buzzmetrics that had a complete overlap with this reference dataset to validate the classification results obtained by our approach.

### 4.5 Effect of Link Polarity

The results in Figure 1 indicate a clear improvement on classifying republican and democratic blogs by applying polar weights to links followed by trust propagation. We get a “cold-start” for democratic blogs and we observe that the overall results are better for republican blogs than democratic blogs. The results being better for republican blogs can be attributed to the observations from [16] that republican blogs typically have a higher connectivity than democratic blogs in the political blogosphere.

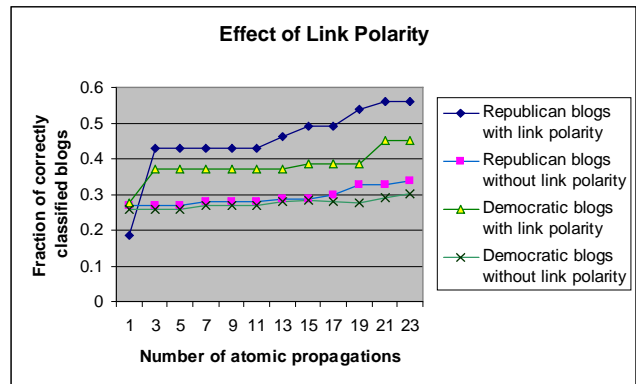


Figure 1: Using polar links for classification yields better results than plain link structure.

We are aware of the fact that the results need to be improved further, however it is interesting to note that there exists an upward swing in the accuracy using polar links. Thus, our idea of using trust propagation to *create* polar links between blogs that do not link to each other directly, helps to classify them. This clearly demonstrates the potential of our approach. We would like to note that the linear curve should not be generalized as a typical characteristic of blogosphere, it might be due to certain attributes of our dataset. We briefly discuss about further analysis of such

trends in the discussion section (section 5).

## 4.6 Effect of Trust propagation

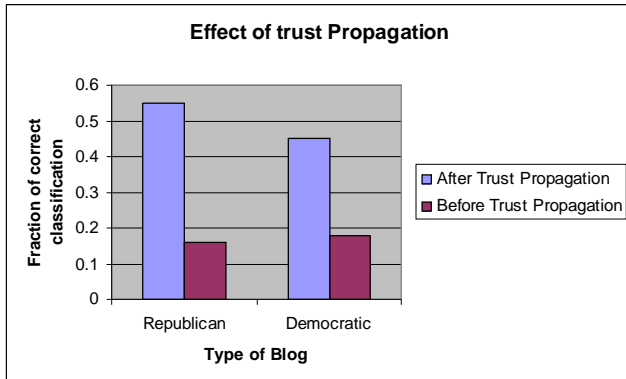


Figure 2: Fraction of correct classification improves significantly after trust propagation *creates* polar links

The results in figure 2 indicate a clear improvement in classification using trust propagation. We can observe upward swing of 40% and 25% in the correct classification of republican and democratic blogs respectively. We believe that these results are promising and can be improved further by using more sophisticated techniques for link polarity analysis.

## 4.7 Effect of window size on polarity determination

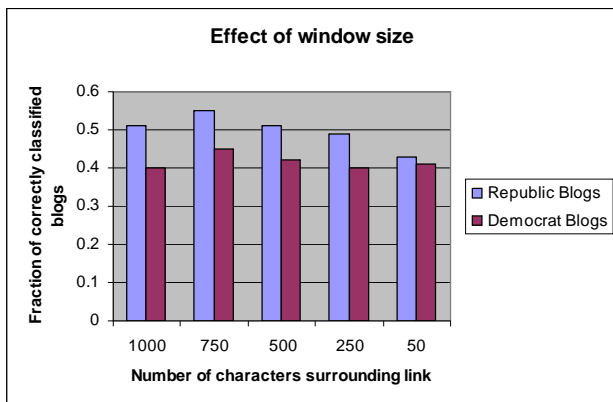


Figure 3: The correctness of classification depends on the optimal window size (around 750 characters) and decays on both sides of the optimal window.

The results in figure 3 indicate that 750 characters was the most appropriate window size for our dataset. If the window size is too small, our system becomes susceptible to short non-opinionated phrases around the link (e.g. *here is what xyz says*) which leads to a zero match of token words to corpus words in text surrounding link. On the other hand, if the window size is too large, our system becomes susceptible to analyzing text unrelated to the opinion expressed around the link. Another source of misinterpretation is the presence of other links in our window. We are evaluating some more heuristics to use more accurate tokens for corpus match (one such heuristic is to stop extending the

window from the link whenever we hit the window size  $x$  or another link).

## 4.8 Sample polarity computations

The table in figure 4 depicts polarity values computed between some pairs of *influential* democratic and republican blogs. We present this data as a quick measure of demonstrating the potential of our work and make the following observations.

1. Trust propagation was effective in predicting the accurate polarity for DK-AT, even though our text processing did not yield the correct polarity initially.
2. Trust propagation retained the sign of polarity if the initial computed sign of polarity was correct (e.g., AT-DK). In fact, trust propagation helped in assigning correct polarities to non-existent links (e.g., AT-IP).
3. The numbers in *italics* indicate the instances where trust propagation failed to assign correct sign to the polarity. However, notice that none of these had any polarity value to start with, so even if trust propagation did not assign the right sign to the link; it helped the clustering process for other blogs by establishing a connection between these blogs. We plan to work on a detailed analysis of such failures in order to get an insight into the effectiveness of our heuristics for link polarity determination. A preliminary analysis indicates that such failures are most likely due to the fact that there are fewer than three links between most blogs in our dataset, hence averaging over such small dataset leads to incorrect sentiment prediction occasionally.

| From-To | Num links | Polarity before trust propagation | Polarity after trust propagation |
|---------|-----------|-----------------------------------|----------------------------------|
| MM-MM   | 0         | N/A                               | +1.007                           |
| MM-DK   | 0         | N/A                               | -9.290                           |
| MM-IP   | 10        | +1.000                            | +1.370                           |
| MM-AT   | 0         | N/A                               | <b>+3.530</b>                    |
| DK-MM   | 0         | N/A                               | -9.290                           |
| DK-DK   | 0         | N/A                               | +8.570                           |
| DK-IP   | 0         | N/A                               | <b>+9.570</b>                    |
| DK-AT   | 20        | -0.084                            | +3.260                           |
| IP-MM   | 8         | +1.000                            | +1.030                           |
| IP-DK   | 6         | +1.000                            | <b>+9.570</b>                    |
| IP-IP   | 0         | N/A                               | +1.060                           |
| IP-AT   | 0         | N/A                               | -3.640                           |
| AT-MM   | 0         | N/A                               | <b>+3.530</b>                    |
| AT-DK   | 5         | 0.342                             | +3.260                           |
| AT-IP   | 0         | N/A                               | -3.640                           |
| AT-AT   | 0         | N/A                               | +1.241                           |

MM -<http://michellemalkin.com>, DK-<http://dailykos.com>  
 IP-<http://instapundit.com>, AT-<http://atrios.blogspot.com>

Figure 4: Polarity values for some influential blogs in our dataset

- We realized the need to enforce a lower bound on the number of sentiment words found in our text analysis before performing link polarity computation. Guha’s model could have worked better if we had set the polarity to zero for all such cases where  $N_p + N_n$  was below two.

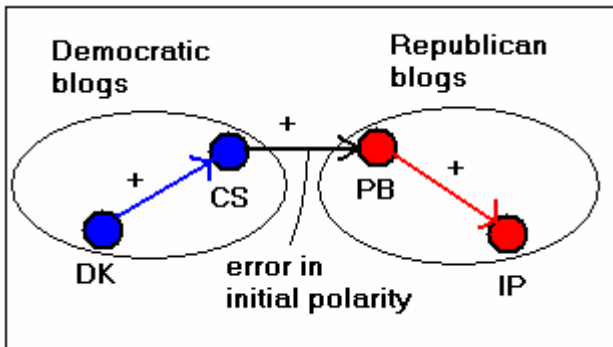


Figure 5: Incorrect initial polarity computation on CS-PB link resulted in positive polar link between DK and IP (+9.570)  
 CS – <http://crooksandliars.com>, PB – <http://powerlineblog.com>

Figure 5 presents a specific case of our analysis. Notice that we computed correct polarity for DK-CS and PB-IP links. However, the text surrounding the link where CS expressed an opinion about PB contained only one positive word “nice” and since we did not use a lower threshold as explained above, we assigned an incorrect value to this link. This incorrect value resulted in establishing a transitive connection between DK and IP after trust propagation leading to the wrong sign for polar link between them.

- Our validation techniques did not involve computing trust score for a blog *foo* from *influential* blogs in both communities. This implies that polar links help us by providing multiple ways to find like-minded blogs for *foo*. Thus, AT – IP polarity can correctly classify AT even if AT – MM polarity is incorrect. However, we are working on finding more sophisticated techniques to perform such validations in graphs having more than two communities and hence, we did not rely on non-scalable methodologies for our validations.

## 5. DISCUSSION

We are aware that we need to analyze results for our approach on a larger dataset. We are also investigating better techniques of validating our results and exploring various heuristics to determine topic of the link. Thus, topic as an extra attribute to the link would give us a fine-grained detail on positive or negative sentiment about a topic over a link and we believe that there are interesting applications of what we would like to term as “topical link polarity”. We are also investigating new clustering techniques that incorporate polarity of links in the distance measure matrix and some of our preliminary results further confirm the effectiveness of link polarity. The idea of using link polarity suits well for all such domains where there exists a distinct set of different opinions (e.g. sports, windows vs. linux etc) and we believe that it has potential for deducing sub-communities from communities as well.

While we are optimistic about our approach, we would like to note that the traditional clustering techniques [18, 19, 20] should be preferred over our approach when the graph is strongly connected. As explained before, the key contribution of our

approach lies in classifying the *marginal* nodes (which either do not link or link very sparingly to the tightly connected cluster nodes). The idea of link polarity can help in predicting the swings in such *marginal* nodes and the temporal analysis of such swings can be very beneficial for advertising applications.

This paper presents preliminary results of our on-going work that demonstrates the effectiveness and feasibility of using polar links as evidence for clustering blogs into communities. Most trust models in use today rely on having *biased* links between nodes and our *polar* links can fit in such models perfectly. The focus of our future work is to make effective use of such polar links in various trust models to determine trustworthy nodes on web graphs.

## 6. CONCLUSION

We describe a novel approach for classifying blogs into predefined sets by applying positive or negative weights to links connecting the blogs. We validated our approach against a labeled dataset and the preliminary results are impressive. We use shallow natural language processing for the text around the links to determine the sentiments of one blog about another. This simple way of sentiment detection augmented by propagating trust using well-known trust models classifies the blogs with high accuracy. The results demonstrate the potential of using polar links for trust determination problems on web graphs and our future work will be focused on addressing this problem.

## 7. ACKNOWLEDGMENTS

We thank Nielson BuzzMetrics for making the blog dataset available. We thank Lada A. Adamic and Natalie Glance for allowing us to use their labeled dataset of democratic and republican blogs for our experimental validations.

## 8. REFERENCES

- Guha R, Kumar R, Raghavan P, Tomkins A. Propagation of trust and distrust. In: *Proceedings of the Thirteenth International World Wide Web Conference*, New York, NY, USA, May 2004. ACM Press, 2004.
- D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW*, pages 491–501, 2004.
- E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, New York, NY, USA, May 2004.
- D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD*, pages 78–87, 2005.
- L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. blogs: Who gets the scoop? In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.
- R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW*, pages 568-576, 2003.

- [7] J. M. Kleinberg. Bursty and hierarchical structure in streams. *Data Min. Knowl. Discov.*, 7(4):373-397,2003.
- [8] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks*, 31(11-16):1481-1493, 1999.
- [9] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW*, 2006.
- [10] K. Shen and L. Wu. Folksonomy as a complex network, Sep 2005.
- [11] C. Marlow. Audience, structure and authority in the weblog community. In *54th Annual Conference of the International Communication Association*, 2004.
- [12] B. Yu and M. P. Singh. A social mechanism of reputation management in electronic communities. In *Cooperative Information Agents*, pages 154–165, 2000.
- [13] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in P2P networks. In *Proceedings of the 12<sup>th</sup> International World Wide Web Conference*, pages 640–651, 2003.
- [14] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in P2P networks. In *Proceedings of the 12<sup>th</sup> International World Wide Web Conference*, pages 640–651, 2003.
- [15] NielsenBuzzmetric, [www.nielsenbuzzmetrics.com](http://www.nielsenbuzzmetrics.com)
- [16] Lada A. Adamic and Natalie Glance, "The political blogosphere and the 2004 US Election", in *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005)*
- [17] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *KDD*, pages 419–428,2005.
- [18] B. Tseng, J. Tatemura, and Y. Wu. Tomographic Clustering To Visualize Blog Communities as Mountain Views. In *Proceedings of the 2<sup>rd</sup> Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 15<sup>th</sup> World Wide Web Conference, May 2005.
- [19] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004.
- [20] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.
- [21] Arun Qamra and Belle Tseng and Edward Y. Chang. "Mining blog stories using community-based and temporal clustering", *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 58--67, 2006